ORIGAMIWORLDBENCH: BENCHMARKING VISUAL INSTRUCTION-CONDITIONED NEXT-STATE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in image editing models have largely centered on natural-language instructions. In real-world diagrams and manuals, however, actions are often specified visually through arrows, lines, and symbols and may induce state transitions of the target object. Generating the resulting state from such visual instructions therefore requires models both to interpret visually specified actions and to faithfully realize their geometric and structural consequences. To study this capability, we introduce **OrigamiWorldBench**, a benchmark for next-state generation conditioned on visually specified actions. Origami is well suited to this setting because origami operations can be visually specified and their outcomes obey clear geometric and structural constraints. Given a current-state image overlaid with a visual prompt and an auxiliary text instruction describing the operation, models directly generate the corresponding one-step-ahead state image. To evaluate geometric and structural fidelity, which is difficult to capture with conventional image-similarity metrics, we introduce an aspect-wise VLM-as-a-Judge protocol. Using OrigamiWorldBench, we evaluate a set of recent image generation and editing models. To quantify the gap between current models and human performance under the same visual instructions, we additionally compare model predictions with origami states physically folded by human participants. We further validate the proposed aspect-wise VLM evaluation against human judgments and conduct an input-cue analysis. We plan to release the benchmark dataset and code upon acceptance.

1 INTRODUCTION

Recent image editing models (Wu et al., 2025; 2026a; Liu et al., 2025) have acquired rich visual priors from large-scale vision-language data and are capable of diverse image transformations following natural-language instructions. However, ***can these models interpret visually specified actions and generate the resulting state?*** In the real world, instructions are not always conveyed through language alone. In diagrams and manuals, visual cues such as arrows, lines, and symbols often indicate where and in which direction an action should be performed.

Moreover, such instructions often involve state transitions of the target object rather than the simple addition or removal of visual elements. As the need to visually predict the consequences of actions in real-world settings grows, generative models are increasingly expected not only to interpret the specified action, but also to predict and generate the resulting visual state. For edits involving such state transitions, evaluating visual quality alone is not sufficient. This raises a second question: ***can models faithfully realize the geometric and structural changes induced by an action?***

Together, these questions highlight two requirements for action-conditioned visual state prediction: interpreting visually specified actions and faithfully realizing their geometric and structural consequences. Yet existing benchmarks rarely evaluate both capabilities within a unified image-editing setting. In particular, the following two limitations remain.

(1) Limited support for visually specified actions. Many existing benchmarks for action-conditioned image generation and editing assume natural-language instructions, although some operations are difficult to specify precisely through text alone. Even when visual prompts are used,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

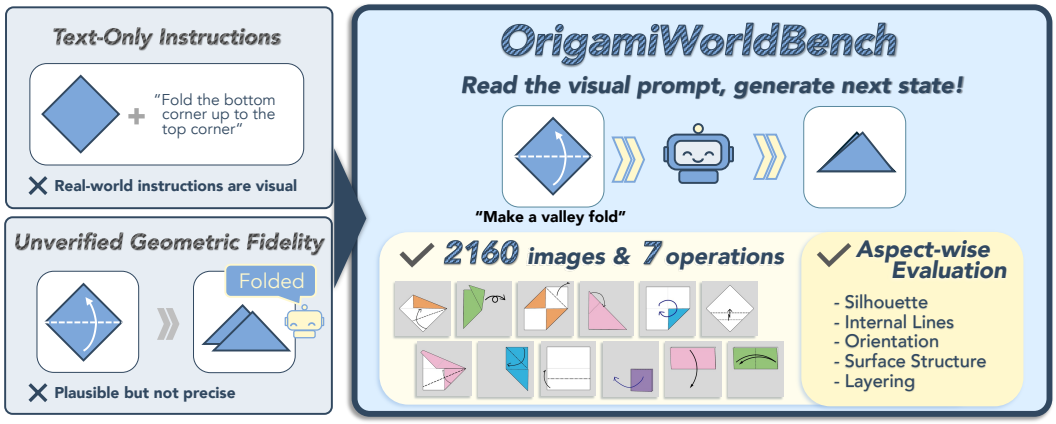


Figure 1: **OrigamiWorldBench** evaluates whether models can ground visually specified origami actions and faithfully generate their geometric and structural consequences. Given a current state with overlaid visual cues and an action instruction, models directly generate the corresponding one-step-ahead state image. The benchmark contains 2,160 samples and evaluates geometric and structural fidelity along five aspects.

they typically indicate predefined trajectories or regions rather than convey the action itself. As a result, these benchmarks do not directly evaluate whether generative models can interpret actions from visual cues and ground them in the current state.

(2) Limited evaluation of geometrically precise editing. Many existing image editing benchmarks primarily evaluate whether the requested semantic or spatial change is reflected in the output, rather than whether precise geometric correspondence with the input is preserved (Zhao et al., 2026b; Li et al., 2026b; Zhu et al., 2026). Meanwhile, many visual reasoning and puzzle benchmarks evaluate geometric transformations through question answering or multiple-choice prediction (Xu et al., 2025; Spencer et al., 2026), where models select among candidate states that provide possible geometric structures rather than constructing the resulting state themselves. These settings therefore do not assess whether models can faithfully realize geometric state transitions in image space.

More recently, generative visual reasoning benchmarks have begun to require models to directly produce reasoning outcomes as images or videos (Liu et al., 2026; Han et al., 2026), while related world-model benchmarks evaluate action-conditioned future-state prediction (Motamed et al., 2026; Xue et al., 2026; Feng et al., 2026; Ye et al., 2026). However, the former primarily focus on latent-rule induction, puzzle solving, or geometric construction, while the latter emphasize physical, temporal, or state consistency. To our knowledge, no existing benchmark systematically evaluates both capabilities in a unified setting: interpreting an action directly from visual cues and generating its one-step consequence with fine-grained geometric and structural fidelity.

To address these challenges, we introduce **OrigamiWorldBench**, a controlled benchmark for geometry-aware visual action-guided editing using origami. Given a current state and a single origami operation, the next origami state is governed by strong geometric and structural constraints. Origami operations are difficult to specify precisely using text alone and are naturally represented in folding diagrams using visual symbols such as arrows and crease lines. These properties make origami suitable for our setting. Origami next-state generation thus allows us to test whether a model merely produces a plausible image or correctly interprets a visual instruction as an action and faithfully realizes its geometric and structural consequences.

Our main contributions are threefold: (i) we construct a new benchmark dataset for directly generating the one-step-ahead origami state from the current origami state and a visually specified origami action, enabling end-to-end evaluation of visual action grounding, action-conditioned state prediction, and structure-preserving image editing; (ii) we introduce an aspect-wise VLM-as-a-Judge (Chen et al., 2024) protocol that evaluates generated states along multiple geometric and structural dimensions; and (iii) using the proposed benchmark, we comprehensively evaluate recent image generation and editing models and report their performance.

2 Related Work

2.1 Visual Action-Guided Image Editing

Recent image editing models have substantially improved their ability to follow natural-language instructions and perform diverse editing operations (Wu et al., 2025; 2026a; Liu et al., 2025). Several benchmarks accordingly evaluate image editing that requires complex instruction understanding and reasoning (Zhao et al., 2026b; Li et al., 2026b; Qian et al., 2025; Wu et al., 2026b). However, most assume that the desired operation is specified primarily through language. Recent work has explored visual cues as editing instructions (Shi et al., 2024; Lu et al., 2024; Geng et al., 2025), alongside benchmarks for visual-instruction-driven editing (Zhang et al., 2026). Visual cues are particularly useful for specifying properties such as location, direction, and shape that are difficult to express precisely through language alone. Related work on visual procedural understanding has also studied mapping symbols in diagrams and manuals to executable actions or plans (Zhang et al., 2023; Wang et al., 2022; Zhang et al., 2025; Li et al., 2026a; Chen et al., 2026). These studies highlight the importance of visual instruction grounding, but typically produce recognition results, plans, or structured action representations rather than visual consequences of the grounded actions. In contrast, OrigamiWorldBench requires models to ground visual cues specifying origami actions and directly generate the resulting next state, extending visual instruction grounding to action-conditioned visual state prediction.

2.2 Geometry-Aware Prediction and Generation

Geometric and structural correctness is an important aspect of spatial reasoning. Existing spatial reasoning benchmarks, including those in origami-based and embodied settings, evaluate folding-induced transformations, crease-pattern understanding, 2D-to-3D reasoning, and next-state prediction (Nanovski et al., 2026; Xu et al., 2025; Spencer et al., 2026; Agarwal et al., 2026; Feng et al., 2026). However, many are formulated as question-answering, multiple-choice, code-generation, or other structured-prediction tasks, without requiring models to construct the inferred next state as an image. More recently, generative visual reasoning and spatially grounded editing benchmarks have begun to evaluate reasoning through direct visual generation (Liu et al., 2026; Zhao et al., 2026a; Wei et al., 2026; Han et al., 2026; Zhu et al., 2026). These tasks, however, mainly focus on latent-rule induction, puzzle solving, or geometric construction rather than one-step state transitions induced by visually specified actions with fine-grained geometric correspondence to the input. Related world-model benchmarks evaluate future-state prediction from observations and actions, emphasizing physical, temporal, or state consistency (Motamed et al., 2026; Xue et al., 2026; Ye et al., 2026; Li et al., 2025). Origami generation has also been studied under geometric constraints, mainly in structured representation or program spaces (Huang et al., 2026; Zahavy et al., 2026). Our work complements these directions by jointly evaluating diagrammatic action grounding, one-step state-transition reasoning, and geometrically faithful next-state generation in pixel space.

3 OrigamiWorldBench

3.1 Task Formulation

We formulate origami next-state generation as an action-conditioned state-transition problem in pixel space. Let S_{current} denote an image of the current folded state, and let S_{next} denote the state image obtained after applying a single origami operation. Rather than specifying the operation using explicit geometric parameters, we provide a *visual prompt* V, consisting of graphical operation marks, together with an auxiliary *action instruction text* T. The visual prompt is overlaid on S_{current} and uses graphical marks such as fold lines and arrows to indicate the location and direction of the operation. The action instruction text T provides no spatial information and specifies only the operation type and a brief description of its meaning (see Figure 2). Instead, the generation model must infer which region of S_{current} should be manipulated and in which direction from the visual prompt. For rotation, T additionally specifies the rotation angle, while the rotation direction must still be inferred from the visual prompt. Let $\mathcal{O}(S, V)$ denote the overlay operator that composes the visual prompt V onto a state image S. The model then takes the overlaid image and the action

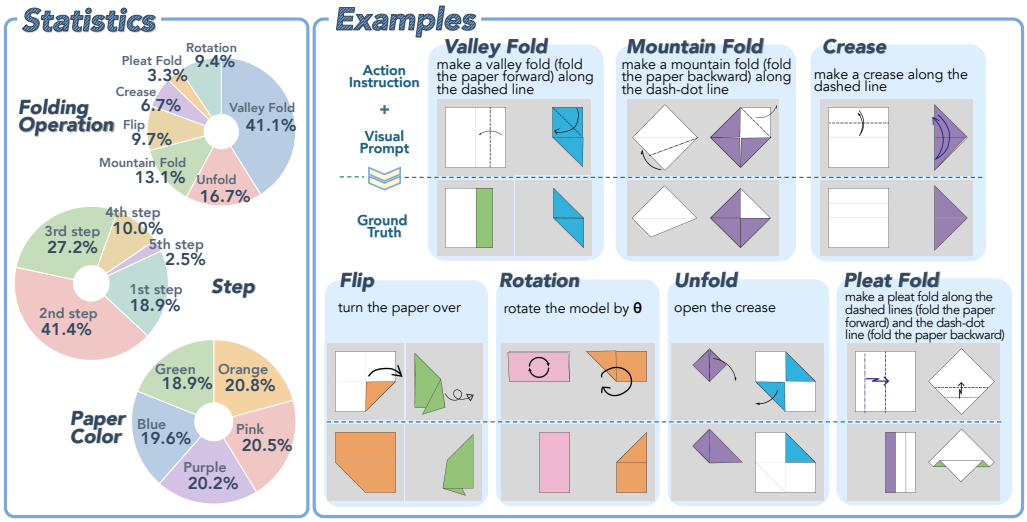


Figure 2: **Overview of OrigamiWorldBench.** Left: sample distributions across operation types, folding steps, and paper colors. Right: representative examples of the seven operation types, showing the action instruction T, visual prompt V, and ground-truth next state S_{next}. Action instructions follow the template "Using the provided image of the origami model, {`operation-specific phrase`} in the direction indicated by the arrow."

instruction text as input:

$$\hat{S}\text{next} = \Phi\left(\mathcal{O}(S\text{current}, V), T\right), \tag{1}$$

where Φ denotes the evaluated model and $\hat{S}_{\text{next}}$ is the predicted next-state image.

The visual prompt consists of two mark layers:

$$V = (v_{\text{line}}, v_{\text{arrow}}), \tag{2}$$

where either layer may be empty when it does not apply to the prescribed operation. When present, the line layer v_{line} specifies the relevant fold or crease location: a dashed line indicates a valley-fold position, whereas a dash–dot line indicates a mountain-fold position. The arrow layer v_{arrow} indicates how the operation should be performed, including its direction and motion pattern.

Solving Eq. 2 therefore requires a model to (i) identify the prescribed operation from the text, (ii) ground the visual prompt in the current state, (iii) infer the resulting state transition, and (iv) render the next state while preserving the paper's geometric and structural properties.

3.2 Data Creation and Dataset Statistics

OrigamiWorldBench is constructed from 63 canonical state images and 78 unique one-step state transitions, expanded into 2,160 samples through controlled state transformations and variations in operation-mark appearance. Figure 2 summarizes the benchmark and shows representative examples of the seven operation types, including the action instruction T, visual prompt V, and ground-truth next state S_{next}. We construct the dataset in three stages: creating canonical state images, designing operation-specific visual marks, and composing samples with controlled transformations and mark variants. We then report statistics on the resulting dataset.

State images. We created all canonical state images as vector graphics in Adobe Illustrator[1] using a unified drawing convention and exported them as raster images with a consistent visual style. The canonical state set was designed to be unique up to 90° rotations and reflections; that is, no two canonical states are related by a multiple of a quarter-turn rotation or by reflection. Rotated and reflected versions were introduced only at the composition stage as controlled transformations.

[1] https://www.adobe.com/products/illustrator.html

Operation marks. For the seven operation types in OrigamiWorldBench—*Valley Fold*, *Unfold*, *Mountain Fold*, *Flip*, *Rotation*, *Crease*, and *Pleat Fold*—we designed graphical operation marks that instantiate the visual prompt V. For each operation type, we prepared six rendering variants that preserve the same operation semantics while varying their visual appearance. These variants follow conventional origami notation and introduce diversity through different arrow geometries, drawing styles, and color schemes.

Composition and mark-variant expansion. Using a dedicated authoring application, we composed the canonical state images and operation marks while applying controlled rotations, reflections, and paper-color variations. The application records the position and orientation of each arrow and fold line, allowing the visual prompt to remain geometrically consistent with the corresponding state transition. This process yields 360 base configurations.

For each base configuration, we render six variants of the operation marks corresponding to the prescribed operation while keeping the underlying current state, next state, action instruction, and operation geometry fixed, resulting in $360 \times 6 = 2{,}160$ samples in total. Thus, the six samples derived from a single base configuration share the same S_{current}, S_{next}, T, and underlying prompt geometry and differ only in the visual rendering of V. This controlled expansion enables us to analyze sensitivity to the appearance of visual instructions independently of the underlying operation and state transition.

Dataset statistics. Figure 2 shows the distribution of samples across the seven operation types. The most frequent operation type is *Valley Fold* (41.1%), followed by *Unfold* (16.7%) and *Mountain Fold* (13.1%). Each sample is also labeled with the folding-step index at which the corresponding transition occurs within its folding sequence, ranging from 1 to 5. Steps 2 and 3 are the most frequent, accounting for 41.4% and 27.2% of the samples, respectively. To introduce visual diversity, paper colors are assigned randomly, resulting in an approximately uniform distribution across the five colors. For evaluation purposes, however, we exclude paper color because, when the reverse side of the paper is not visible in the current state, its color cannot be inferred from the input and therefore cannot be expected to be generated correctly.

3.3 Evaluation Metrics

We employ a VLM-as-a-Judge (Chen et al., 2024) protocol to evaluate the generated images. For each sample, we provide the generated image and its corresponding ground-truth image to the vision language model (VLM) and ask it to independently assess five criteria of geometric and structural fidelity: *overall silhouette*, *internal lines*, *orientation*, *surface structure*, and *visible layering*. Each criterion is scored as a binary yes-or-no judgment, corresponding to 1 or 0, respectively. We then average the five binary scores to obtain a sample-level score on a 0–1 scale.

We adopt this aspect-wise evaluation rather than asking the VLM to assign a single holistic correctness score for two reasons. First, origami states can exhibit highly complex geometry and fine-grained structural details. When a VLM is asked to make a single overall judgment, subtle discrepancies may be overlooked. By evaluating each structural aspect independently, we constrain the judge to focus on specific properties of the generated state and thereby reduce the risk of missing such errors. Second, a holistic scalar score can be sensitive to variation across repeated VLM judgments, particularly for partially correct outputs that fall between clearly correct and clearly incorrect cases. In contrast, decomposing the evaluation into independent binary decisions provides a more constrained scoring procedure before aggregation. We examine the effect of this aspect-wise evaluation strategy in detail in Section 4.4.

4 Experiments

4.1 Experimental Setup

We evaluate recent image generation and editing models. For proprietary models, we evaluate `gpt-image-[1.5/2]` (OpenAI, 2025) and `Gemini-3.1-Flash-Image` (OpenAI, 2026a). For open-source models, we evaluate `FLUX.1-Kontext-dev` (Labs et al., 2025), `FLUX.2-klein-[4B/9B]` (Labs, 2025), `FLUX.2-dev` (Labs, 2025), `OmniGen2` (Wu et al.,

Table 1: Overall benchmark results on all 2,160 samples. The best score in each column within each model group is shown in **bold**.

Model	Silhouette	Lines	Orientation	Surface	Layering	Average
Proprietary Models						
GPT-Image-2	**0.420**	**0.286**	**0.669**	**0.188**	**0.290**	**0.371**
GPT-Image-1.5	0.276	0.142	0.389	0.063	0.179	0.210
Gemini-3.1-Flash-Image	0.335	0.143	0.470	0.112	0.192	0.250
Open-Source Models						
FLUX.1-Kontext-dev	**0.243**	**0.189**	**0.457**	0.100	**0.149**	**0.228**
FLUX.2-klein-4B	0.213	0.120	0.414	0.081	0.114	0.189
FLUX.2-klein-9B	0.225	0.093	0.411	0.071	0.115	0.183
FLUX.2-dev	0.170	0.047	0.158	0.032	0.085	0.099
OmniGen2	0.234	0.165	0.320	0.090	0.143	0.190
Qwen-Image-Edit-2511	0.175	0.063	0.405	**0.112**	0.127	0.176
JoyAI-Image-Edit	0.172	0.057	0.316	0.046	0.070	0.132
LongCat-Image-Edit	0.126	0.024	0.219	0.033	0.052	0.091

Table 2: Performance by operation type and folding-step index. The macro averages are the unweighted means across the seven operation types and five step indices, respectively. The best score in each column within each model group is shown in **bold**.

	Operation Type								Folding Step					
Model	Valley Fold	Unfold	Mountain Fold	Flip	Rotation	Crease	Pleat Fold	Macro Avg.	Step 1	Step 2	Step 3	Step 4	Step 5	Macro Avg.
Proprietary Models														
GPT-Image-2	**0.256**	**0.325**	**0.315**	**0.296**	**0.735**	**0.931**	**0.303**	**0.451**	**0.408**	**0.422**	**0.286**	**0.282**	**0.507**	**0.381**
GPT-Image-1.5	0.173	0.165	0.226	0.207	0.303	0.344	0.300	0.245	0.295	0.217	0.142	0.200	0.252	0.221
Gemini-3.1-Flash-Image	0.222	0.194	0.278	0.189	0.435	0.354	0.222	0.271	0.336	0.257	0.187	0.217	0.352	0.270
Open-Source Models														
FLUX.1-Kontext-dev	**0.134**	**0.193**	**0.245**	0.180	0.444	0.685	**0.097**	**0.283**	**0.213**	**0.264**	**0.191**	0.215	0.163	**0.209**
FLUX.2-klein-4B	0.120	0.183	0.137	0.160	0.265	**0.747**	0.011	0.232	0.128	0.226	0.162	0.188	**0.267**	0.194
FLUX.2-klein-9B	0.096	0.178	0.126	**0.293**	0.226	0.708	0.000	0.233	0.107	0.224	0.164	0.184	0.230	0.182
FLUX.2-dev	0.075	0.074	0.055	0.147	0.282	0.107	0.003	0.106	0.052	0.065	0.129	0.202	0.256	0.141
OmniGen2	0.117	0.174	0.182	0.152	0.395	0.533	0.050	0.229	0.123	0.222	0.168	**0.238**	0.178	0.186
Qwen-Image-Edit-2511	0.095	0.074	0.121	0.047	**0.589**	0.700	0.067	0.242	0.156	0.196	0.148	0.188	0.263	0.190
JoyAI-Image-Edit	0.088	0.064	0.143	0.102	0.399	0.288	0.006	0.156	0.088	0.117	0.142	0.222	0.244	0.163
LongCat-Image-Edit	0.063	0.087	0.057	0.028	0.129	0.406	0.022	0.113	0.066	0.107	0.078	0.095	0.096	0.089

2026a), `Qwen-Image-Edit-2511` (Wu et al., 2025), `JoyAI-Image-Edit` (Song et al., 2026), and `LongCat-Image-Edit` (Team et al., 2025). We use the default inference parameters for each model. We use `Gemini-3.7-Flash` (OpenAI, 2026b) as the VLM-based evaluator. Details of the evaluated models and prompting setup are provided in Appendix A.

4.2 Benchmark Results

Overall performance. Table 1 reports the overall results on all 2,160 samples. GPT-Image-2 achieves an average score of 0.371, while FLUX.1-Kontext-dev obtains the highest average among the evaluated open-source models at 0.228. Across models, orientation is generally preserved more successfully than fine-grained structural properties such as surface structure and visible layering. The overall scores remain substantially below 1.0, highlighting the difficulty of generating structurally faithful next states from visually specified origami actions.

Performance by operation type and folding step. Table 2 reports performance by operation type and folding-step index. For some models, differences emerge across operation types. In particular, *Rotation* and *Crease* yield relatively high scores for several models, whereas *Valley Fold* and *Pleat Fold* remain more challenging, suggesting that next-state prediction performance depends substantially on the type of state transition. Performance also tends to decrease from earlier to later folding steps for several models, particularly from Steps 1–2 to Steps 3–4, suggesting that structurally more complex states can be more difficult to predict. This trend is not monotonic, however, and several

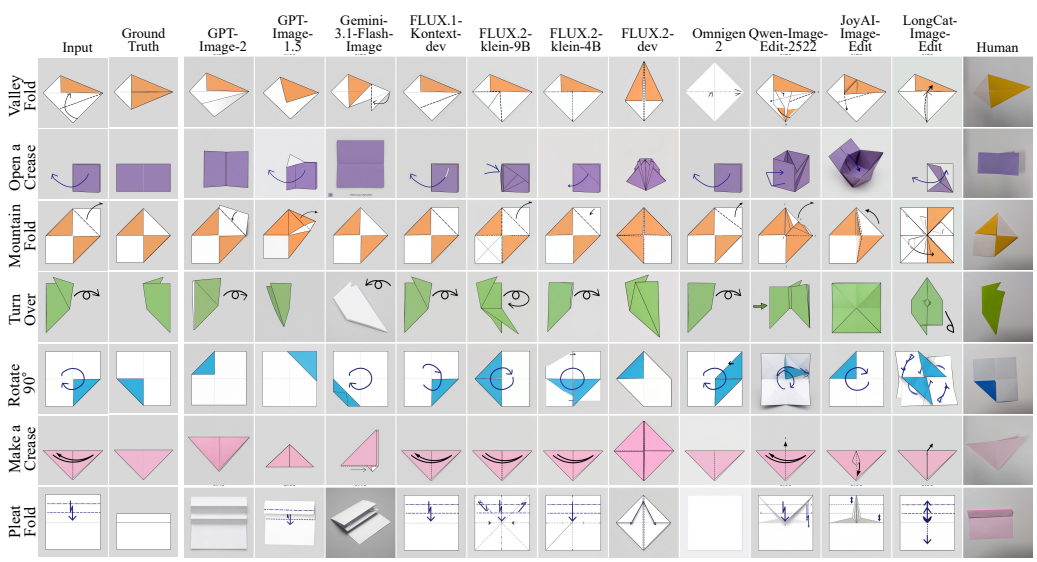


Figure 3: Qualitative results.

models achieve higher scores at Step 5. Because Step 5 contains comparatively few samples, these estimates are likely to be more variable and should therefore be interpreted with caution.

Qualitative results. Figure 3 shows outputs from the evaluated models. A key challenge is to transform only the relevant regions while preserving the unaffected geometry. All models exhibit failures in this respect, with shape distortions particularly common among several open-source models. Even for a simple 90° rotation, some models alter the origami geometry rather than preserving it under rotation. In the *Unfold* example, where the visible area should approximately double, only GPT-Image-2 correctly captures the expected transformation among the examples shown.

For the analyses in the following subsections, we construct a configuration-level subset to keep evaluations involving human participants practically feasible while preserving coverage of all base configurations. As described in Section 3, each of the 360 base configurations is expanded into six samples that differ only in the rendering of the operation marks. We uniformly sample one variant from each configuration, yielding a 360-sample analysis subset, which is the smallest configuration-level subset that retains one representative of every base configuration. To quantify sampling variability, we repeated the sampling procedure 1,000 times for each model and measured the variation in the resulting 360-sample averages. Across the 11 models, the mean half-width of the central 95% interval was approximately ± 0.024, indicating limited variation due to subsampling.

4.3 Human-Folded Reference

To assess whether the visual prompts and action instructions provide sufficient information to infer the intended origami transitions, we asked human participants to physically perform the corresponding operations using the same task information provided to the models. If humans can consistently infer and execute the intended transitions from the visual prompts and action instructions, this provides evidence that the task specification itself is sufficiently informative and that the performance gap primarily reflects limitations of current generative models. We therefore use the human-folded results as a human reference upper bound.

Human folding protocol. The participants consisted of eight individuals, including both male and female participants, ranging from their 20s to 60s and with varying levels of prior origami experience. Each participant was assigned approximately 45 samples, collectively covering all 360 samples in the analysis subset. For each sample, participants were provided with a physical sheet of paper folded into the current state, together with the visual prompt V and action instruction T. No additional information about the target state was provided, ensuring that participants performed the

Table 3: Comparison with the human-folded reference on the 360-sample analysis subset.

Model	Silhouette	Lines	Orientation	Surface	Layering	Average
Unedited Input	0.332	0.297	0.370	0.108	0.228	0.267
Human-Folded	0.942	0.822	0.914	0.856	0.919	0.891
Proprietary Models						
GPT-Image-2	**0.436**	**0.281**	**0.650**	**0.186**	**0.278**	**0.366**
GPT-Image-1.5	0.297	0.158	0.375	0.069	0.167	0.213
Gemini-3.1-Flash-Image	0.331	0.156	0.489	0.111	0.181	0.253
Open-Source Models						
FLUX.1-Kontext-dev	**0.242**	**0.183**	**0.447**	0.108	**0.161**	**0.228**
FLUX.2-klein-4B	0.225	0.119	0.417	0.075	0.100	0.187
FLUX.2-klein-9B	0.208	0.069	0.394	0.061	0.092	0.165
FLUX.2-dev	0.158	0.044	0.147	0.031	0.089	0.094
OmniGen2	0.231	0.164	0.325	0.089	0.144	0.191
Qwen-Image-Edit-2511	0.172	0.058	0.431	**0.111**	0.128	0.180
JoyAI-Image-Edit	0.181	0.061	0.317	0.053	0.078	0.138
LongCat-Image-Edit	0.117	0.025	0.231	0.028	0.053	0.091

Table 4: Correlation between human and VLM evaluations.

	Silhouette	Lines	Orientation	Surface	Layering	Average
Pearson r	0.708	0.657	0.609	0.640	0.652	0.802
Spearman ρ	0.675	0.578	0.614	0.586	0.602	0.773

task using the same visual and textual instructions available to the models. Participants performed the operation that they interpreted as being instructed and placed the paper at the specified position and in the orientation they considered correct. The resulting states were photographed, center-cropped to a square, and resized to 1024×1024. Further details are provided in Appendix A.4.

Human–model performance comparison. The human-folded reference achieves an average score of 0.891, substantially higher than all evaluated generative models. In comparison, GPT-Image-2, the strongest model on this subset, obtains 0.366. This large gap suggests that the visual instructions contain sufficient information for humans to infer and execute the intended transitions, while current generative models still struggle to translate the same instructions into structurally faithful next-state images. Notably, most models perform worse than the unedited-input baseline, suggesting that the editing process often introduces substantial distortions of the origami structure.

4.4 Validation of VLM-based Evaluation

Our evaluation framework relies on a VLM to assess the structural fidelity of generated next states across five criteria. To validate this evaluation, we conduct a human study on 360 images by uniformly sampling 45 images from each of eight prediction sources, consisting of seven generative models and the human-folded images described in Section 4.3. We compare these human judgments with the VLM evaluations and further examine whether aspect-wise evaluation yields higher correlation with human judgments than a single holistic judgment.

User study design. We recruited 35 participants through Prolific[2], screened only for English fluency, and compensated them at an hourly rate of £14.58; the median completion time was approximately 36 minutes. Each participant evaluated 60 randomly assigned images, with assignments distributed evenly, resulting in 3–7 evaluations per image (5.83 on average). Participants independently assigned binary scores using the same five-criterion rubric as the VLM.

Correlation with human evaluation. Table 4 reports the correlation between human and VLM evaluations. Across the 360 evaluated images, the average scores yield a Pearson correlation of

[2] https://www.prolific.com/

Table 5: Input-cue analysis for GPT-Image-2 and Gemini-3.1-Flash-Image. A checkmark indicates that the corresponding cue is provided, whereas a cross mark indicates that it is removed.

Models	Condition			Silhouette	Lines	Orientation	Surface	Layering	Average
	Arrow	Line	Action						
GPT-Image-2	✓	✓	✓	0.436	**0.281**	0.650	**0.186**	0.278	0.366
	✗	✓	✓	**0.447**	0.247	**0.683**	**0.186**	**0.300**	**0.373**
	✓	✗	✓	0.364	0.214	0.489	0.131	0.228	0.285
	✗	✗	✓	0.375	0.247	0.358	0.164	0.244	0.278
	✓	✓	✗	0.356	0.167	0.450	0.114	0.211	0.259
Gemini-3.1-Flash-Image	✓	✓	✓	0.331	0.156	0.489	0.111	0.181	0.253
	✗	✓	✓	**0.342**	**0.164**	**0.528**	**0.119**	**0.233**	**0.277**
	✓	✗	✓	0.319	0.128	0.344	0.097	0.169	0.212
	✗	✗	✓	0.239	0.114	0.258	0.081	0.142	0.167
	✓	✓	✗	0.306	0.114	0.406	0.094	0.194	0.223

$r = 0.802$ and a Spearman correlation of $\rho = 0.773$. At the individual-criterion level, Pearson correlations range from 0.609 to 0.708, showing consistently positive correlations across all five aspects. The mean inter-rater variance in the average score was 0.104.

Effect of aspect-wise evaluation. We further compare our aspect-wise evaluation with a holistic evaluation in which the VLM directly assigns a single overall judgment. The aspect-wise aggregation yields higher correlation with human judgments, increasing the Pearson correlation from 0.725 to 0.802 and the Spearman correlation from 0.674 to 0.773. This result supports our choice to decompose structural fidelity into independently evaluated aspects before aggregation.

4.5 Input-Cue Analysis

For next-state prediction, the action instruction provides a macro-level description of the operation, while the visual prompt provides detailed spatial information about how it should be performed. The task is therefore designed to require the model to combine semantic information from the action instruction with spatial information from the visual prompt. If removing a cue leaves performance unchanged or improves it, this may indicate that the model does not effectively use or correctly interpret that information. Table 5 reports the input-cue analysis for GPT-Image-2 and Gemini-3.1-Flash-Image, the two highest-scoring models in our benchmark. We remove the arrow, line, and action instruction cues to examine their contributions to next-state prediction. Removing only the arrow does not degrade performance and instead improves it for both models. In contrast, removing either the line or the action instruction consistently reduces performance, while removing both visual cues causes a substantial drop. These results suggest that the models make limited use of the directional information encoded by the arrow and rely more heavily on the spatial constraint provided by the line and the semantic information in the action instruction.

5 Conclusion

We introduced **OrigamiWorldBench**, a new benchmark for evaluating next-state generation conditioned on visually specified actions. Unlike many existing image editing and visual reasoning benchmarks, OrigamiWorldBench jointly evaluates visual action grounding, state-transition reasoning, and faithful next-state generation. We also introduced an aspect-wise VLM-as-a-Judge protocol for evaluating geometric and structural fidelity. Our evaluation of recent image editing models revealed substantial challenges in accurately realizing state transitions and preserving fine-grained structure. Human participants performing the same operations substantially outperformed current models, highlighting a large gap in translating visual instructions into correct next states. The proposed aspect-wise evaluation showed strong correlation with human judgments, while our input-cue analysis indicated limited use of directional information encoded by arrows. Together, these results show that grounding visually specified actions and faithfully generating their geometric and structural consequences remain major challenges. OrigamiWorldBench provides a testbed for measuring progress toward visual action understanding and structurally faithful next-state generation.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, assisting with translation, and supporting qualitative and thematic data analysis. Generative AI models were also used as systems under evaluation and as evaluators, as described in the relevant sections of the paper. We did not use generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning or reformatting datasets, or interpreting results. Tasks related to formulating mathematical claims or assisting with their proofs were not applicable to this work. Additionally, we used generative AI tools for summarizing and analyzing existing literature, sourcing and searching for information, improving the readability of the manuscript, identifying relevant literature, and proposing a title for the paper. We reviewed all AI-assisted work. The authors manually verified each cited reference against its source and reviewed and tested all AI-assisted code to confirm its correctness and intended behavior. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The benchmark dataset includes only common and basic origami operations, and all benchmark samples were created by the authors. Therefore, the dataset does not reproduce specific copyrighted origami works or instructions. The benchmark dataset itself does not contain personally identifiable information or sensitive private data. Human participants were involved only in the human-folded reference and evaluation studies described in Section 4.3, Section 4.4, Appendix A.4, and Appendix A.5. We do not identify any substantial risk of harm to specific individuals or broader adverse societal impacts arising from the use of this benchmark.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. Details of the experimental setup, including the evaluated model versions and prompting templates, are provided in Appendix A.

REFERENCES

Naaisha Agarwal, Yihan Wu, Yichang Jian, Yikuan Hu, Nishad Mansoor, Mohan Li, Yifei Peng, Wang-Zhou Dai, Yao-Xiang Ding, and Emanuele Sansone. Origamibench: An interactive environment to synthesize flat-foldable origamis. *arXiv preprint arXiv:2603.13856*, 2026.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024.

Guanhua Chen, Yutong Yao, Shenghe Sun, Ci-jun Gao, Shudong Liu, Lidia S. Chao, Feng Wan, and Derek F. Wong. Chain-of-procedure: Hierarchical visual-language reasoning for procedural QA. In *Findings of the ACL*, 2026.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

ZhiYuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shi Shubin, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, Rushuai Yang, Ruichuan An, Leqi Zheng, Weijie Wang, Shuang Chen, Sicheng Xu, Yaobo Liang, Jiaolong Yang, and Baining Guo. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. In *ICLR*, 2026.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *CVPR*, 2025.

Haonan Han, Jiancheng Huang, Xiaopeng Sun, Junyan He, Rui Yang, Jie Hu, Xiaojiang Peng, Lin Ma, Xiaoming Wei, and Xiu Li. Vigor-bench: How far are visual generative models from zero-shot visual reasoners? *arXiv preprint arXiv:2603.25823*, 2026.

Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Trans. Inf. Theory*, 8:179–187, 1962.

Yanjia Huang, Yunuo Chen, Ying Jiang, Jinru Han, Zhengzhong Tu, Yin Yang, and Chenfanfu Jiang. Learn2fold: Structured origami generation with world model planning. *arXiv preprint arXiv:2603.29585*, 2026.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence, 2025. URL https://bfl.ai/blog/flux-2. Accessed: 2026-09-25.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Danrui Li, Jiahao Zhang, Bernhard Egger, Moitreya Chatterjee, Suhas Lohit, Tim K. Marks, and Anoop Cherian. Assemblybench: Physics-aware assembly of complex industrial objects. *arXiv preprint arXiv:2605.12845*, 2026a.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. GIR-bench: Versatile benchmark for generating images with reasoning. In *ICLR*, 2026b.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yutong Liu, Nan Huang, Xu Cao, and James M. Rehg. Thinking in pictures: A systematic benchmark for reasoning-driven image generation. *arXiv preprint arXiv:2609.02864*, 2026.

Jingyi Lu, Xinghui Li, and Kai Han. Regiondrag: Fast region-based image editing with diffusion models. In *ECCV*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *WACV*, 2026.

Stiliyan Nanovski, Mrinal Dhume, and Rafael Bidarra. Difficulty-based generation of paperfolding puzzles. In *FDG*, 2026.

OpenAI. The new chatgpt images is here, 2025. URL https://openai.com/index/new-chatgpt-images-is-here/. Accessed: 2026-09-25.

OpenAI. Gemini 3.1 flash image, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-image/. Accessed: 2026-09-25.

OpenAI. Gemini 3.7 flash, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed: 2026-09-25.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. ISSN 2835-8856.

Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. GIE-Bench: Towards Grounded Evaluation for Text-Guided Image Editing. *arXiv preprint arXiv:2505.11493*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. *ICCV*, 2011.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, 2024.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Ryan Spencer, Roey Yaari, Ritvik Vemavarapu, Joyce Yang, Steven Ngo, and Utkarsh Sharma. Gamibench: Evaluating spatial reasoning and 2d-to-3d planning capabilities of MLLMs with origami folding tasks. In *AAAI*, 2026.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Ruocheng Wang, Yunzhi Zhang, Jiayuan Mao, Chin-Yi Cheng, and Jiajun Wu. Translating a visual lego manual to a machine-executable plan. In *ECCV*, 2022.

Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13:600–612, 2004.

Jingxuan Wei, Caijun Jia, Xi Bai, Xinglong Xu, Siyuan Li, Linzhuang Sun, Bihui Yu, Conghui He, Lijun Wu, and Cheng Tan. Ggbench: A geometric generative reasoning benchmark for unified multimodal models. In *CVPR*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ziyi Xia, Ze Liu, Chaofan Li, Haoge Deng, Kun Luo, Bo Zhang, Jiajun Zhang, Dong Liu, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. In *CVPR*, 2026a.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang YU, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. KRIS-bench: Benchmarking next-level intelligent image editing models. In *NeurIPS*, 2026b.

Rui Xu, Dakuan Lu, Zicheng Zhao, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, and Xu Yinghui. ORIGAMISPACE: Benchmarking multimodal LLMs in multi-step spatial reasoning with mathematical constraints. In *NeurIPS*, 2025.

Haotian Xue, Yipu Chen, Liqian Ma, Zelin Zhao, Lama Moukheiber, Yuchen Zhu, and Yongxin Chen. Acwm-phys: Investigating generalized physical interaction in action-conditioned video world models. *arXiv preprint arXiv:2605.08567*, 2026.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Tom Zahavy, Shaobo Hou, Thomas Tumiel, James Doran, Francesco Faccio, Xidong Feng, Alex Havrilla, Igor Khytryi, Chenglei Li, Lisa Schut, Vivek Veeriah, Arijan Abrashi, Michał Kosmulski, Robert J. Lang, Nick Robinson, Brandon Wong, Marcus Chiam, Gloria Fang, and Satinder Singh. Corigami: An ai pipeline for co-designing flat-foldable visually recognisable origami. *arXiv preprint arXiv:2606.26299*, 2026.

Huanyu Zhang, Xuehai Bai, Chengzu Li, Chen Liang, Haochen Tian, Haodong Li, Ruichuan An, Yifan Zhang, Anna Korhonen, Zhang Zhang, Liang Wang, and Tieniu Tan. How well do models follow visual instructions? vibe: A systematic benchmark for visual instruction-driven image editing. *arXiv preprint arXiv:2602.01851*, 2026.

Jiahao Zhang, Anoop Cherian, Yanbin Liu, Yizhak Ben-Shabat, Cristian Rodriguez, and Stephen Gould. Aligning step-by-step instructional diagrams to video demonstrations. In *CVPR*, 2023.

Jiahao Zhang, Anoop Cherian, Cristian Rodriguez, Weijian Deng, and Stephen Gould. Manual-pa: Learning 3d part assembly from instruction diagrams. In *ICCV*, 2025.

Changpeng Zhao, Yiren Song, and Jinpeng Wang. Genpuzzle: Benchmarking visual reasoning in image generation models. *arXiv preprint arXiv:2609.05902*, 2026a.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *NeurIPS*, 2026b.

Muzhi Zhu, Shunyao Jiang, Huanyi Zheng, Zekai Luo, Hao Zhong, Anzhou Li, Kaijun Wang, Jintao Rong, Yang Liu, Hao Chen, Tao Lin, and Chunhua Shen. Exploring spatial intelligence from a generative perspective. In *CVPR*, 2026.

A Experimental Setup Details

A.1 Model Setup

This section lists the inference settings used for each model. Parameters not mentioned below are left at the default values of the corresponding API or Diffusers[3] pipeline.

Gemini 3.1 Flash Image. We use `gemini-3.1-flash-image` through the official API with resolution $= 512$ and the default temperature (1.0).

GPT-Image Series. We use `gpt-image-1.5-2025-12-16` and `gpt-image-2-2026-04-21` through the official API with quality set to medium and an output size of 1024×1024. Since the API does not expose a random seed, the outputs are not reproducible at the pixel level.

Qwen-Image-Edit. We use `Qwen-Image-Edit-2511`[4] with 30 inference steps, a true CFG scale of 4.0, a guidance scale of 1.0, and a negative prompt consisting of a single space. The seed is fixed to 0.

FLUX Series. We evaluate four FLUX variants: `FLUX.2-klein-4B`[5] and `FLUX.2-klein-9B`[6] with 4 inference steps and a guidance scale of 1.0; `FLUX.2-dev`[7]

[3] https://github.com/huggingface/diffusers
[4] https://huggingface.co/Qwen/Qwen-Image-Edit-2511
[5] https://huggingface.co/black-forest-labs/FLUX.2-klein-4B
[6] https://huggingface.co/black-forest-labs/FLUX.2-klein-9B
[7] https://huggingface.co/black-forest-labs/FLUX.2-dev

with 28 inference steps and a guidance scale of 4.0; and `FLUX.1-Kontext-dev`[8] with 28 inference steps and a guidance scale of 2.5. The output resolution is set to the input resolution, with each dimension rounded down to the nearest multiple of 16. The seed is fixed to 0.

OmniGen2. We use `OmniGen2`[9] with 50 inference steps and the Euler scheduler. The text and image guidance scales are set to 5.0 and 2.0, respectively, and classifier-free guidance is applied over the entire denoising trajectory, with a CFG range of $[0.0, 1.0]$. We use the default negative prompt from the official implementation and a maximum text sequence length of 1024. The output size is 1024×1024, and the seed is fixed to 0.

JoyAI-Image-Edit. We use `JoyAI-Image-Edit-Diffusers`[10] with 40 inference steps and a guidance scale of 4.0, without a negative prompt. The seed is fixed to 0. This model supports only single-image input.

LongCat-Image-Edit. We use `LongCat-Image-Edit`[11] with 50 inference steps, a guidance scale of 4.5, and an empty negative prompt. Following the official README, the seed is fixed to 43.

A.2 Action Instruction Text Template

For all models evaluated in our benchmark, we use a common set of text prompts. For each operation type, the action instruction provides only a simple description of the action and does not provide the spatial information required to perform the intended operation without the visual prompt, which consists of cues such as arrows and fold lines. In Table 6, we provide the operation-specific phrases used for the default setting and each condition in the input-cue analysis. When a visual cue is removed, the corresponding reference is also removed from the text instruction so that the instruction does not explicitly refer to a cue that is absent from the input image. For rotation, θ is instantiated as 90°, 180°, or 270° according to the sample.

[8] https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev
[9] https://huggingface.co/OmniGen2/OmniGen2
[10] https://huggingface.co/jdopensource/JoyAI-Image-Edit-Diffusers
[11] https://huggingface.co/meituan-longcat/LongCat-Image-Edit

Table 6: Operation-specific editing instructions used in the input-cue analysis. The common prefix "Using the provided image of the origami model," is omitted from the table. Checkmarks and crosses indicate whether each cue is retained, when applicable, or removed, respectively.

Input Cue	Default Setting ✓ Arrow ✓ Line ✓ Action	w/o Arrow ✗ Arrow ✓ Line ✓ Action	w/o Line ✓ Arrow ✗ Line ✓ Action	w/o Arrow & Line ✗ Arrow ✗ Line ✓ Action	w/o Action ✓ Arrow ✓ Line ✗ Action
Valley fold	make a valley fold (fold the paper forward) along the dashed line in the direction indicated by the arrow	make a valley fold (fold the paper forward) along the dashed line	make a valley fold (fold the paper forward) in the direction indicated by the arrow	make a valley fold (fold the paper forward)	fold the paper according to the marks shown in the image
Unfold	open the crease indicated in the direction indicated by the arrow	open the crease	open the crease indicated in the direction indicated by the arrow	open the crease	fold the paper according to the marks shown in the image
Mountain fold	make a mountain fold (fold the paper backward) along the dash-dot line in the direction indicated by the arrow	make a mountain fold (fold the paper backward) along the dash-dot line	make a mountain fold (fold the paper backward) in the direction indicated by the arrow	make a mountain fold (fold the paper backward)	fold the paper according to the marks shown in the image
Flip	turn the paper over in the direction indicated by the arrow	turn the paper over	turn the paper over in the direction indicated by the arrow	turn the paper over	fold the paper according to the marks shown in the image
Crease	make a crease along the dashed line as indicated by the arrow	make a crease along the dashed line	make a crease as indicated by the arrow	make a crease	fold the paper according to the marks shown in the image
Pleat fold	make a pleat fold along the dashed lines (fold the paper forward) and the dash-dot line (fold the paper backward) in the direction indicated by the arrows	make a pleat fold along the dashed lines (fold the paper forward) and the dash-dot line (fold the paper backward)	make a pleat fold in the direction indicated by the arrows	make a pleat fold	fold the paper according to the marks shown in the image
Rotation	rotate the model by θ in the direction indicated by the arrow	rotate the model by θ	rotate the model by θ in the direction indicated by the arrow	rotate the model by θ	fold the paper according to the marks shown in the image

A.3 Evaluation Prompt Template

For the VLM-as-a-Judge evaluation, the ground-truth and generated images are resized to 512×512 using Lanczos resampling and arranged side by side in a single panel, with the ground truth on the left and the generated image on the right. In the aspect-wise evaluation, the five criteria (*i.e.*, silhouette, internal lines, orientation, surface structure, and visible layering) are evaluated independently using binary Yes/No judgments. We use a common system prompt across all evaluation criteria to ensure consistent comparison rules and specify which factors should be ignored during evaluation. Specifically, the evaluator is instructed to disregard image scale, paper color, rendering style, and instructional overlays such as arrows and lines, and to compare generated outputs with a 3D appearance or perspective against the top-down ground-truth view by mentally projecting them to the corresponding layout. For each criterion, a separate user prompt instructs the evaluator to judge only the target aspect without considering the remaining criteria. The evaluator is also required to provide a brief rationale and a Yes/No score for each judgment.

In addition, we prepared a prompt for holistic evaluation for comparison with the aspect-wise evaluation. The holistic evaluation follows the same comparison rules but assesses the overall agreement between the generated image and the ground truth using a single score ranging from 0 to 1. The holistic evaluation prompt is derived from the prompt used for our benchmark evaluation, with only the minimal modifications necessary to accommodate this change in evaluation format. Below, we provide the system and user prompt templates used for these evaluations.

System prompt (shared across aspect-wise evaluations)

```
# Role:
You are an objective and precise evaluator of origami simulation
    results.
# Task:
You receive one panel: ground truth (left, labeled GT) and
    candidate (right, labeled Pred).
Decide whether the candidate matches the ground truth ONLY in the
    aspect provided, following the rules below.
# Rules:
Apply the following when comparing:
1. The ground truth is a top-down flat reference. Match the
    candidate to that structure; do not require the candidate
    itself to be drawn as a flat diagram.
2. If the candidate is 3D or has perspective, mentally flatten /
    project it to the same top-down view before comparing.
3. Ignore scale, the specific colors of the paper, and rendering
    style (photo vs illustration). Note that a front/back side
    difference is not a mere color difference.
4. Ignore distorted lines from paper sagging / warping in
    photorealistic depictions.
5. Ignore overlay marks such as arrows, dotted lines, and other
    instructional symbols; do not ignore lines that belong to the
    paper itself (creases and paper edges).
```

User prompt template: Silhouette

```
Decide Yes or No for the aspect below only:
- Silhouette: the coarse outer contour only, ignoring rotation and
    internal details. Ignore slight differences in fold-back angle
    or flap width.
Do not judge other aspects (internal lines, orientation, surface
    structure, visible layering) unless the aspect above explicitly
    says so.

Answer with a binary decision:
```

```
- Yes = the candidate matches the ground truth on this aspect
- No = the candidate does not match on this aspect

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <Yes or No>
```

User prompt template: Internal lines

```
Decide Yes or No for the aspect below only:
- Internal lines: the position and arrangement of all visible
    lines inside the outer contour in the flattened finished model,
    ignoring rotation. Answer Yes only if these lines clearly
    correspond to those in the ground truth. Answer No if any line
    is missing, extra, or misplaced, or if the correspondence is
    uncertain.
Do not judge other aspects (overall silhouette, orientation,
    surface structure, visible layering) unless the aspect above
    explicitly says so.

Answer with a binary decision:
- Yes = the candidate matches the ground truth on this aspect
- No = the candidate does not match on this aspect

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <Yes or No>
```

User prompt template: Orientation

```
Decide Yes or No for the aspect below only:
- Orientation: the in-plane rotation matches the ground-truth
    top-down layout. If the orientation cannot be determined
    because the folds are malformed or entirely different from the
    ground truth, answer No.
Do not judge other aspects (overall silhouette, internal lines,
    surface structure, visible layering) unless the aspect above
    explicitly says so.

Answer with a binary decision:
- Yes = the candidate matches the ground truth on this aspect
- No = the candidate does not match on this aspect

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <Yes or No>
```

User prompt template: Surface structure

```
Decide Yes or No for the aspect below only:
- Surface structure: the arrangement and shape of the visible
    surfaces, and which regions show the front vs back side of the
    paper (regardless of the actual colors), ignoring rotation.
Do not judge other aspects (overall silhouette, internal lines,
    orientation, visible layering) unless the aspect above
    explicitly says so.
```

```
Answer with a binary decision:
- Yes = the candidate matches the ground truth on this aspect
- No = the candidate does not match on this aspect

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <Yes or No>
```

User prompt template: Visible layering

```
Decide Yes or No for the aspect below only:
- Visible layering: the stacking order of overlapping flaps (which
    parts lie on top of which), ignoring rotation.
Do not judge other aspects (overall silhouette, internal lines,
    orientation, surface structure) unless the aspect above
    explicitly says so.

Answer with a binary decision:
- Yes = the candidate matches the ground truth on this aspect
- No = the candidate does not match on this aspect

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <Yes or No>
```

System prompt for holistic evaluation

```
# Role:
You are an objective and precise evaluator of origami simulation
    results.
# Task:
You receive one panel: ground truth (left, labeled GT) and
    candidate (right, labeled Pred).
Decide whether the candidate matches the ground truth ONLY in the
    aspect provided, following the rules below.
# Rules:
Apply the following when comparing:
1. The ground truth is a top-down flat reference. Match the
    candidate to that structure; do not require the candidate
    itself to be drawn as a flat diagram.
2. If the candidate is 3D or has perspective, mentally flatten /
    project it to the same top-down view before comparing.
3. Ignore scale, the specific colors of the paper, and rendering
    style (photo vs illustration). Note that a front/back side
    difference is not a mere color difference.
4. Ignore distorted lines from paper sagging / warping in
    photorealistic depictions.
5. Ignore overlay marks such as arrows, dotted lines, and other
    instructional symbols; do not ignore lines that belong to the
    paper itself (creases and paper edges).
```

User prompt template for holistic evaluation

```
Decide how well the candidate matches the ground truth on this
    single aspect:
```

```
- match: whether the candidate matches the ground truth.

Answer with one score from 0 to 1:
- 1 = the candidate matches the ground truth
- 0 = the candidate does not match the ground truth
- Use intermediate values (for example 0.25, 0.5, 0.75) when the
    match is partial.

Output format (reply in exactly this format):
Reasoning: <brief justification>
Score: <number from 0 to 1>
```

A.4 Details of Human-Folded Data Collection

Participant recruitment and consent. Participants were recruited on a voluntary basis from among the authors' acquaintances. Before beginning the study, participants were informed of the following:

- The purpose of the study and that the collected human-folded results would be used as a reference for comparison with model-generated outputs.
- Participants would manipulate a physical sheet of origami paper according to the displayed instructions, and the resulting state would be photographed.
- The study was not intended to assess participants' individual abilities.
- Participation was voluntary, and participants could decline to participate or withdraw at any time without penalty or disadvantage.
- The experiment could be stopped immediately if participants experienced discomfort or felt unwell.
- The collected results could be reported in academic publications or released as part of a fully anonymized dataset. Separately, participants could optionally consent to being acknowledged by name in the paper upon acceptance. This was not required for participation in the study.

Participants provided consent after reviewing this information and before proceeding with the study.

Inter-participant variability. To assess inter-participant variability, all eight participants also performed the same 13 samples. The resulting folded states were evaluated using the same VLM-based protocol. Participant-level mean scores ranged from 0.862 to 0.954 (SD $= 0.034$), with no significant systematic differences across participants (Friedman test, $\chi^2(7) = 5.82$, $p = 0.56$). These results suggest that participant-specific variability was limited at the aggregate level.

A.5 Human Evaluation

Participant recruitment and consent. Participants were recruited through Prolific. Before beginning the study, participants were informed of the following:

- The purpose and procedure of the study, including that they would evaluate AI-generated origami results using five predefined binary criteria.
- The study involved minimal foreseeable risk, with possible mild visual fatigue or discomfort from viewing images.
- Participation was voluntary, and participants could withdraw at any time without penalty or disadvantage.
- Responses were associated with Prolific IDs for submission management, while no direct personal identifiers were collected.
- The collected data would be retained for five years after the research period, and participants could request deletion of their responses after withdrawal.

- The collected results could be reported in academic publications, and fully anonymized data could be shared or released for scientific research.
- Contact information for the research team was provided for questions regarding the study or collected data.

Participants provided consent after reviewing this information and before proceeding with the study.

Evaluation interface. Figure 4 shows an example of the evaluation interface presented to participants during the human study. The evaluation instructions were adapted from the VLM prompts, with the wording simplified for human participants while preserving the same evaluation criteria and considerations.

B Comparison with Conventional Image-Similarity Metrics

We further compare our VLM-based evaluation with conventional reference-based image-similarity metrics, including CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), HOG (Dalal & Triggs, 2005), Hu Moments (I1) (Hu, 1962), ORB (Rublee et al., 2011), and SSIM (Wang et al., 2004). For each sample, we compute the human overall score by averaging the scores across human raters, and measure its Pearson and Spearman correlations with each automatic evaluation metric. As shown in Table 7, our aspect-wise VLM-as-a-Judge evaluation shows substantially higher agreement with human judgments than the conventional image-similarity metrics. These results indicate that directly assessing the geometric and structural properties of the generated state provides an evaluation that more closely reflects human judgments than generic image-similarity measures.

Table 7: Sample-level correlation between human overall scores and automatic evaluation metrics. ↑ indicates higher is better, while ↓ indicates lower is better.

	CLIP ↑	DINOv2 ↑	HOG ↑	Hu Moments (I1) ↓	ORB ↑	SSIM ↑	**Ours** ↑
Pearson r	0.364	0.249	0.453	-0.128	0.133	0.420	**0.802**
Spearman ρ	0.398	0.279	0.462	-0.332	0.160	0.422	**0.773**

C Benchmark Dataset Samples

In addition to the examples shown in the main text, we present several dataset samples for each operation type in Figure 5. The dataset includes images generated by applying different operations, visual prompts, rotations, reflections, and paper colors to the same underlying state images, with no duplicate images included.

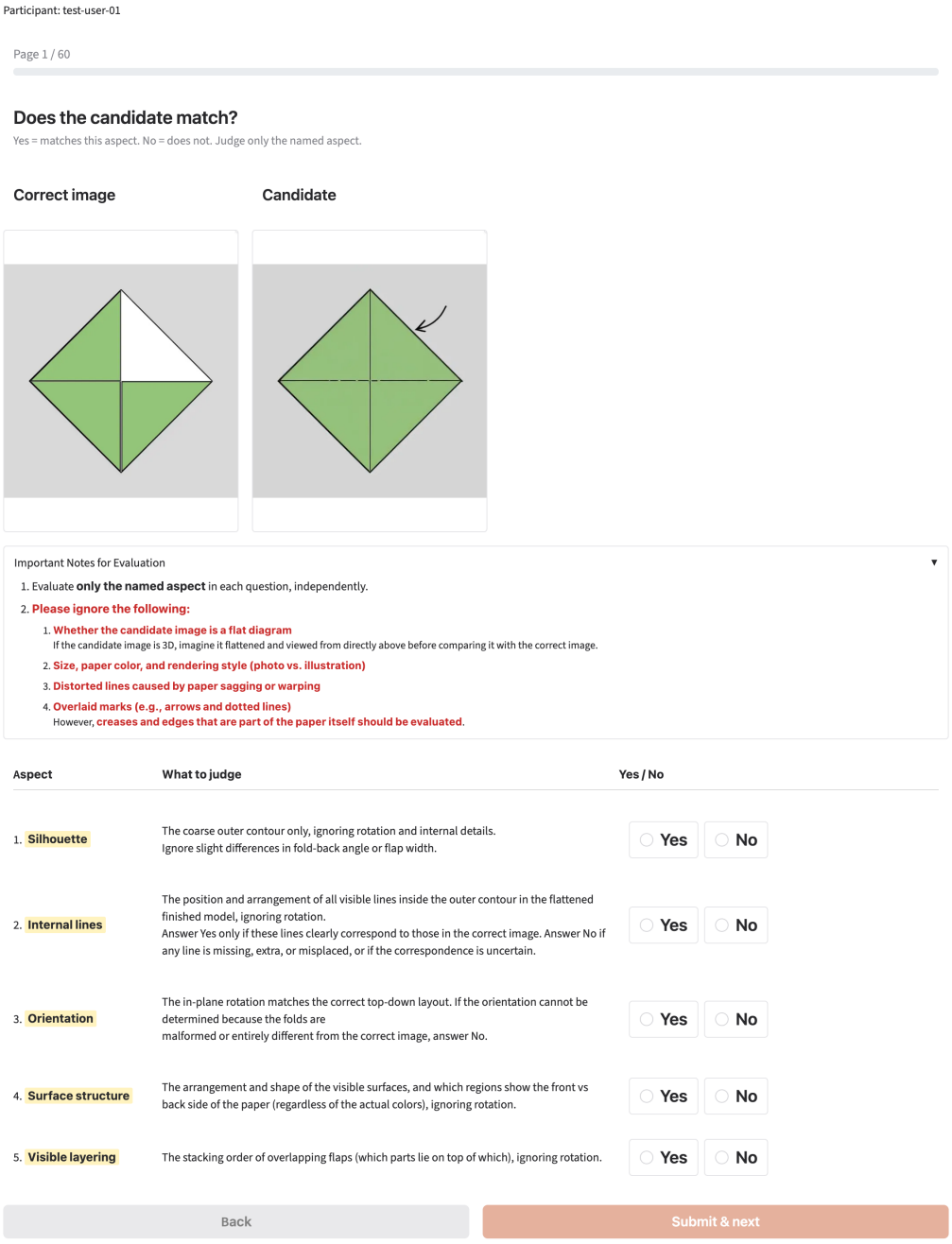


Figure 4: Sample page of the human evaluation interface.

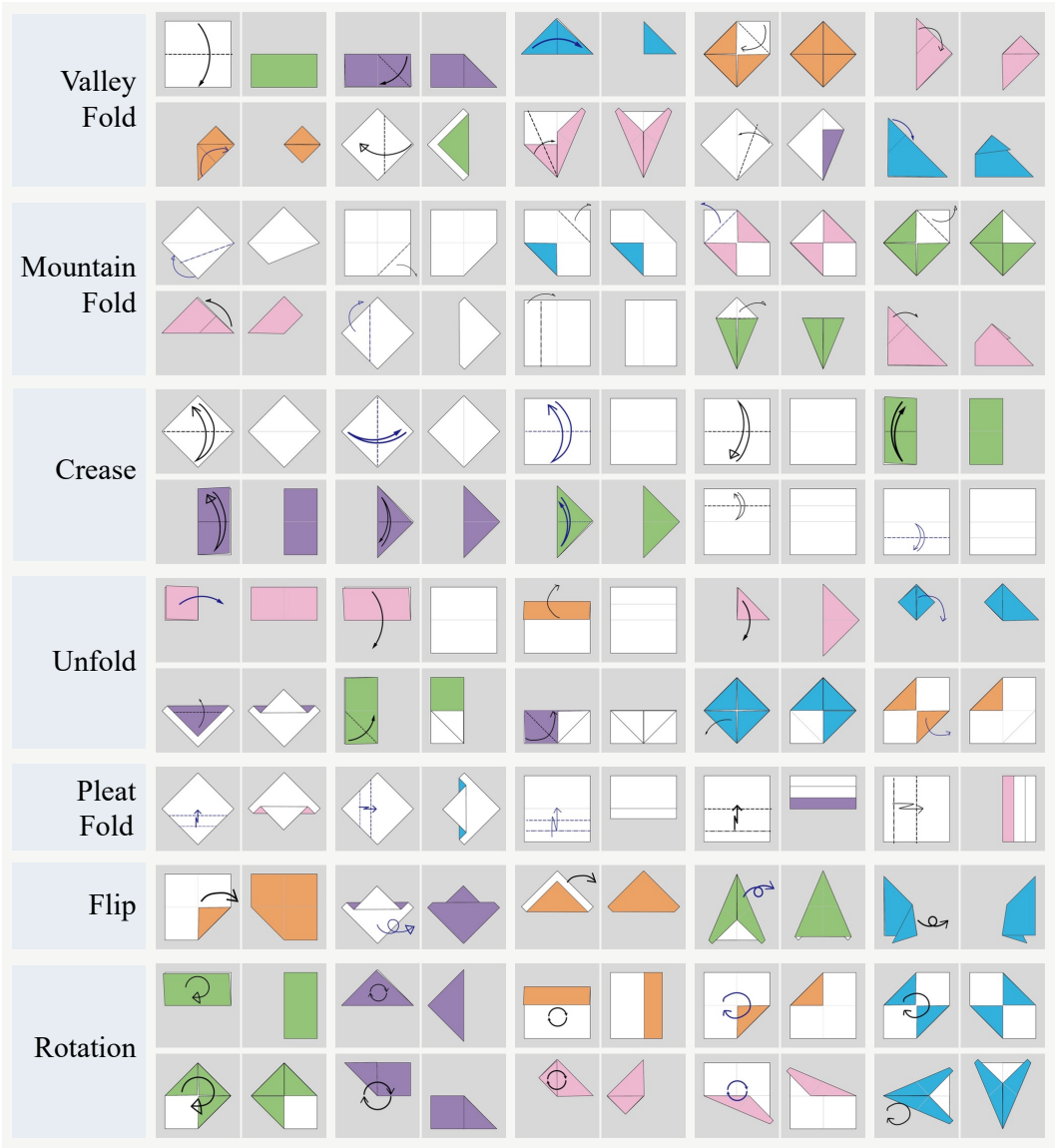


Figure 5: **Additional dataset samples.** For each example, the input image is shown on the left and the corresponding ground-truth next state on the right.