COPY-TRANSFORM-PASTE: Zero-Shot Object-Object Alignment Guided by Vision-Language and Geometric Constraints

Supplementary Material

9. Benchmark Construction

Two curated benchmarks. We release two OOA variants: **(i) Rigid** (translation+rotation only; no scale) and **(ii) Scale-enabled** (translation+rotation+isotropic scale). This separation enables fair comparison to methods that do not support scale.

Benchmark overview. The benchmark consists of 50 mesh–prompt pairs spanning several types of spatial relations, including *on-top* (22), *inside* (9), *held/worn* (9), *insertion/cutting* (6), and *riding/attached* (4).

The object pairs cover diverse categories, including food (14), tools (8), containers (7), characters/humans (10), animals (6), and miscellaneous objects (5). Prompts were written manually to describe a single spatial interaction between the two objects.

Source assets and reference pose. We collected pairs of meshes from Sketchfab. For each pair, we manually placed the objects in their intended relational configuration (e.g., candle inside candlestick). This configuration serves as the reference pose defining the intended relative arrangement between the objects. It fixes only the relative arrangement; meshes are not necessarily canonicalized or upright.

Perturbation protocol (initializations). Starting from the reference pose, we create randomized starting positions by selecting one object at random and:

- **Translate** each mesh independently by a random offset $\Delta \sim \mathcal{U}\big(-10\,\mathbf{L},\, 10\,\mathbf{L}\big)$, where $\mathbf{L} = (L_x, L_y, L_z)$ are the side lengths of its axis-aligned bounding box.
- **Rotate** it by independent Euler angles within $\pm 180°$.
- **Scale** (for the *scale-enabled* benchmark only) isotropically, sampled uniformly from $[0.01,\, 100.0]$.

All draws are uniform and independent; the *rigid* benchmark keeps the scale fixed at 1.0.

Outputs. For each benchmark instance we also provide: (1) the *final aligned meshes*, and (2) the *text prompt*.

10. LLM prompt design for hyperparameter selection

In Sec. 3.6, we described how an LLM is used to estimate three hyperparameters from the object names and scene description. Here, we provide the exact prompts used to query the model. Each prompt is written in natural language and instructs the model to output a single JSON value on one line.

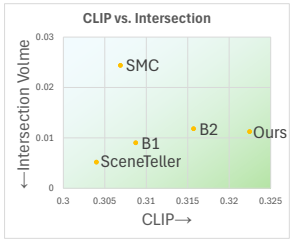
(1) Initial scale. Estimates the real-world relative size between the two objects: `Estimate the relative scale needed so that object1="{object1}" and object2="{object2}" fit together naturally in the desired alignment "{wanted_alignment}". Define size_ratio = bbox_size(object1) / bbox_size(object2). Output exactly one JSON object: {"size_ratio": <float between 0.01 and 100.0>}.`

(2) Penetration policy. Determines whether the final configuration should involve one object penetrating another (e.g., cutting, slicing): `Decide whether achieving alignment "{wanted_alignment}" between object1="{object1}" and object2="{object2} REQUIRES solid-to-solid penetration. Output exactly one JSON object: {"penetration": <true|false>}.`

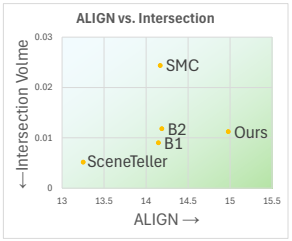
(3) Attachment ratio. Estimates how much surface contact is expected between the two objects in the final arrangement: `For the desired alignment "{wanted_alignment}", estimate the fraction of surface contact between object1="{object1}" and object2="{object2}". Define contact_ratio in [0,1], where 0 = almost no contact and 1 = full surface contact. Output exactly one JSON object: {"contact_ratio": <float 0..1>}.`

Implementation details. We use an instruction-tuned LLaMA-3 model to estimate the required hyperparameters. Each query is repeated up to 10 times, and the first valid JSON-parsed output is used. If no valid output is produced, default parameter values are applied. All prompts are fixed and rely only on the object names and the textual scene description.

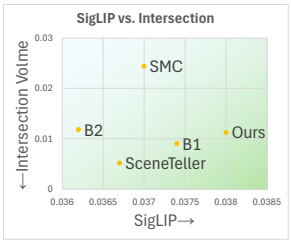
Usage. At test time, we query the LLM with these three templates using only the object names and textual prompt, without any rendered images. The model outputs the three values used to set: (i) the source-object initial scale, (ii) whether the penetration loss is enabled, and (iii) the attachment ratio r controlling the soft-ICP vertex fraction.



(a) CLIP vs. intersection



(b) ALIGN vs. intersection



(c) SigLIP vs. intersection

Figure 13. **Extended trade-off plots.** Vision–language alignment score versus intersection volume. Lower intersection and higher alignment (down-right) is better.

11. User Study Protocol

We include here the full user study protocol used in our evaluation, consisting of the instructions shown to participants (Fig. 16) and an example trial interface (Fig. 17). These materials were presented exactly as shown to all participants during the study.

12. Additional Trade-off Visualizations

Figure 13 provides the full trade-off visualization referenced in the main paper, including results for SigLIP in addition to CLIP and ALIGN.

13. Additional Baseline Comparison and Component Analysis

Additional baseline and ablation comparison. After the initial submission of this work, the implementation of the OOR-diffusion baseline became publicly available. The released resources include a reduced training setup with 9 training objects, since the original training dataset and pretrained weights are not available. As a result, the behavior of the released model differs noticeably from the results reported in the original OOR-diffusion paper, where the method was trained on a substantially larger dataset.

Using the released implementation, we conducted an additional comparison between OOR-diffusion and our method under the same evaluation protocol. In practice, the released model frequently produces configurations in which objects remain spatially separated or loosely aligned, with little direct attachment between them.

Separately, we also evaluate an ablation that replaces the proposed fractional soft-ICP term with standard soft-ICP while keeping all other components unchanged. This experiment isolates the contribution of localized attachment within our framework. For fairness, all methods are evaluated under identical conditions using the same benchmark instances, prompts, and random seeds, with only the alignment method changed.

Method	CLIP ↑	ALIGN ↑	SigLIP ↑	Inter. ↓	Text–Asset ↑	3D Plaus. ↑	Text–Geom. ↑	Overall ↑
Soft-ICP	0.307	13.92	**0.042**	0.022	991.58	988.06	988.06	991.58
OOR-diffusion	0.308	13.91	0.038	0.011	1004.21	1008.26	1008.26	1000.11
Ours (CLIP+Frac. Soft-ICP)	**0.322**	**14.99**	0.041	**0.009**	**1009.18**	**1013.24**	**1013.24**	**1010.04**

Table 5. **Alignment comparison.** Soft-ICP abaltion, OOR-diffusion baseline, and our method across semantic alignment (higher is better), physical plausibility (intersection; lower is better), and VLM-based scores (higher is better). OOR-diffusion results are obtained using the publicly released implementation trained on the available subset of 9 objects.

Figure 14. **Extreme scale diagnostic.** Fixed *sundae* + *cherry* pair with identical meshes, prompt, and seed, varying only the initial cherry-to-sundae volume ratio (shown below each image).

Tab. 5 summarizes the results. Our full method achieves the best CLIP, ALIGN, and VLM-based evaluation scores while also obtaining the lowest intersection values, indicating improved semantic alignment and physical plausibility. The soft-ICP ablation shows weaker attachment behavior, highlighting the benefit of the proposed fractional soft-ICP formulation for encouraging localized and semantically consistent contact between objects.

14. Robustness to Extreme Scale Differences

We additionally evaluate robustness to large size mismatches between objects. In this diagnostic experiment we fix a *sundae–cherry* pair and vary only the initial scale of the cherry across several orders of magnitude while keeping the meshes, prompt, and random seed identical.

Figure 14 shows that correct placement is maintained across a wide range of ratios, with failures appearing only at extremely small volume ratios ($\sim 1 \times 10^{-5}$). This suggests that the optimization procedure remains stable even under substantial scale discrepancies between objects.

15. Analysis of the Text Guidance Signal

To illustrate the discriminative ability of the text guidance signal, we render multiple candidate poses for a fixed object pair and compare their scores under CLIP.

Figure 15 shows a representative *kettle–lid* example. CLIP assigns the highest score to the correct, physically

Figure 15. Candidate lid poses rendered from varying viewpoints under the prompt *"kettle with its lid on top"*.

Policy	CLIP ↑	ALIGN ↑	SigLIP ↑	Intersection ↓
LLM flag (on/off)	0.283	12.30	0.038	0.009
Fixed margin ($c_{\text{pen}} > 0$)	0.283	12.26	0.038	0.010

Table 6. **Penetration policy comparison.** Mean scores on five penetration-intended pairs; both variants yield comparable results.

plausible placement where the lid rests properly on the kettle, while tilted or floating alternatives receive lower scores. This observation suggests that, despite its global formulation, CLIP provides a useful directional signal for optimizing object–object alignment.

16. Penetration Policy Analysis

Our method uses a binary penetration policy inferred from the prompt to determine whether interpenetration between objects is semantically expected (*e.g.*, cutting or insertion scenarios).

We compare this policy against a margin-based variant that always enables the penetration loss. Table Tab. 6 shows that both strategies produce nearly identical results on penetration-expected cases, indicating that the simpler binary policy is sufficient in practice.

17. Additional Image-to-3D Alignment Results

Beyond text-conditioned alignment, our method can also be guided by reference images. Figure 18 presents additional examples of this image-to-3D setting. Each row shows the two input meshes, the guiding reference image, and the final optimized arrangement.

18. Additional Qualitative Comparisons

We include additional per-mesh qualitative comparisons between our method and the four baselines (B1, B2, SceneTeller, and SMC). For each prompt, the figures show (left to right) the target mesh alone followed by the final aligned meshes rendered from four different viewpoints. These supplementary examples cover a diverse set of physical and semantic relationships and illustrate consistent improvements of our method across cases. See Figs. 19 to 26 for the full set of comparisons.

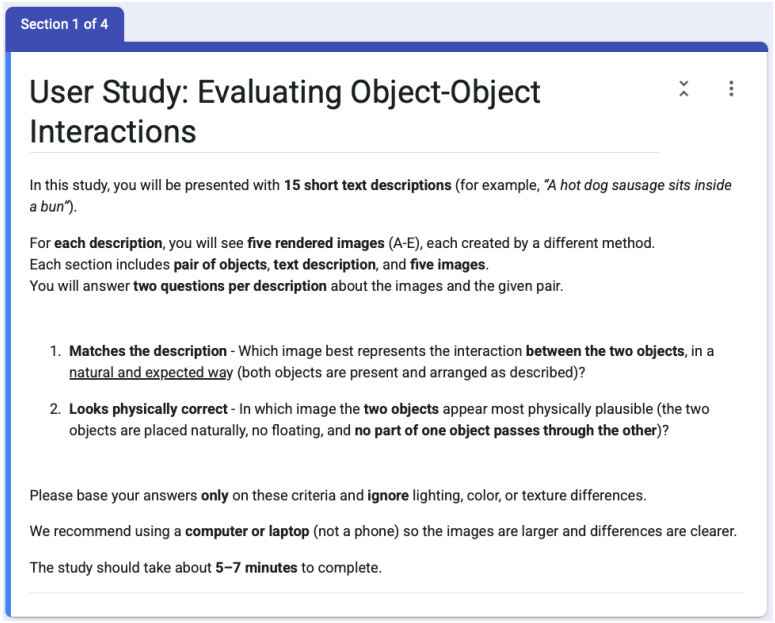


Figure 16. **User study guidelines (screenshot).**

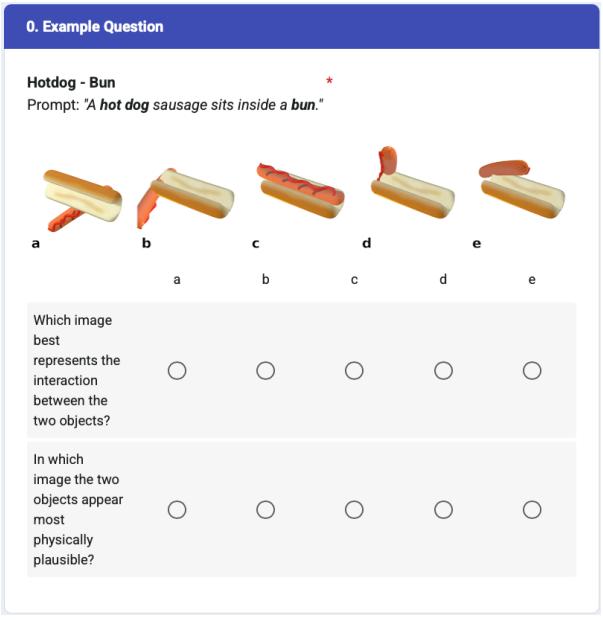


Figure 17. **Example trial (screenshot).**

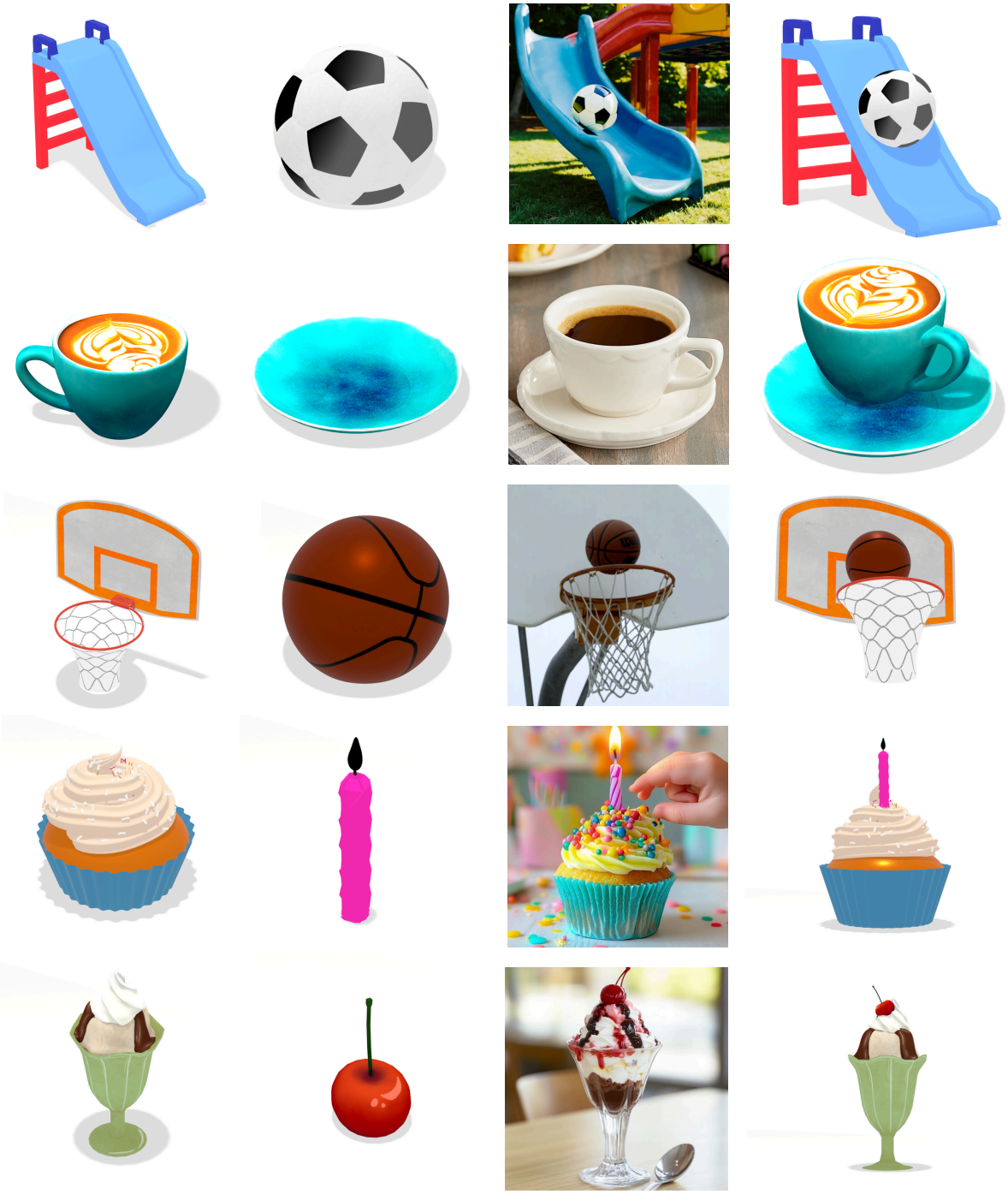


Figure 18. **Additional image-to-3D alignment results.** Each row shows an example of our image-guided alignment process: the two input meshes (left), the reference image used for guidance (middle), and the final optimized placement (right).

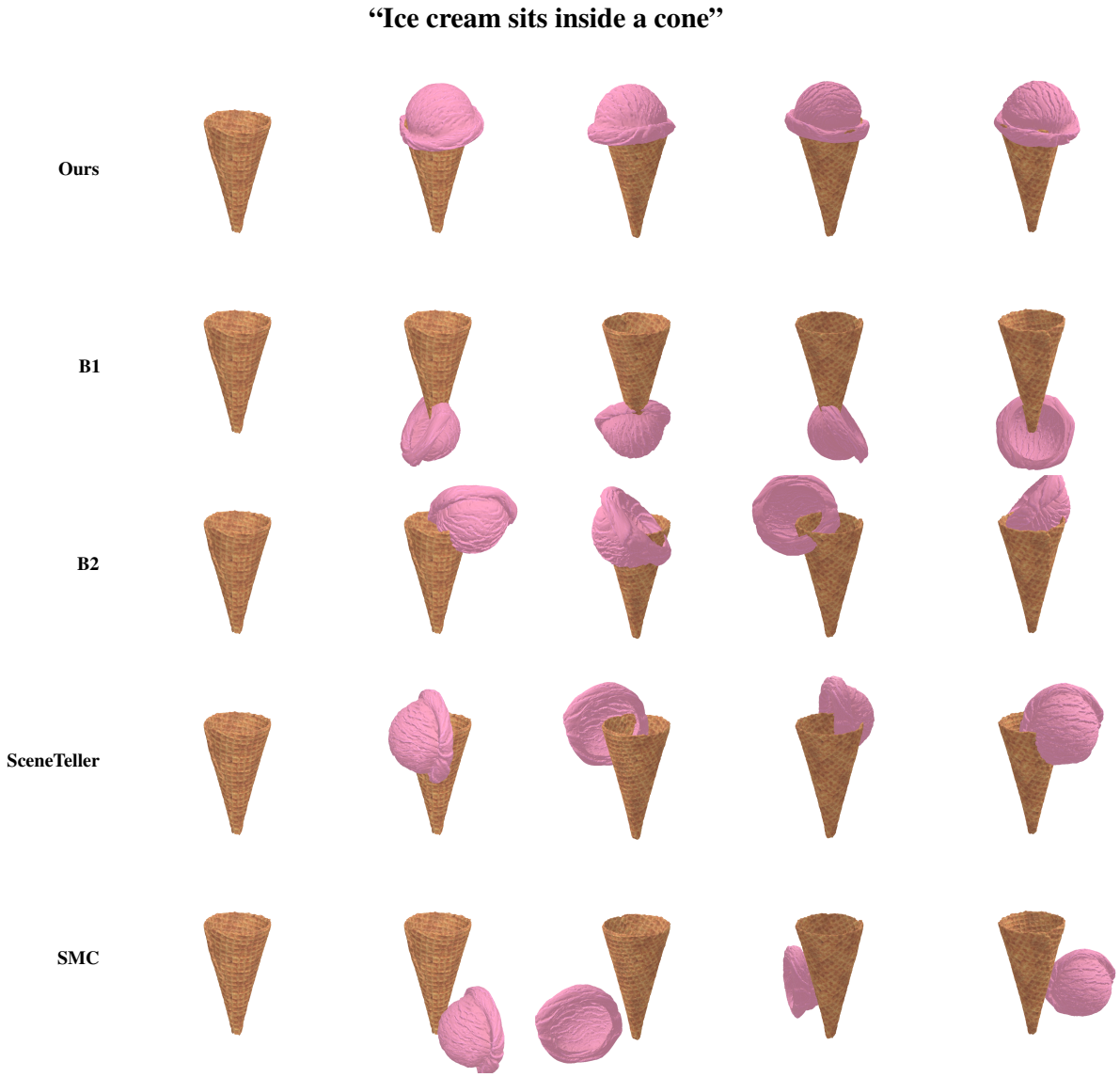


Figure 19. **Per-mesh qualitative comparison: ice cream and cone.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Only our method achieves a text-aligned configuration where the ice cream sits physically inside the cone.

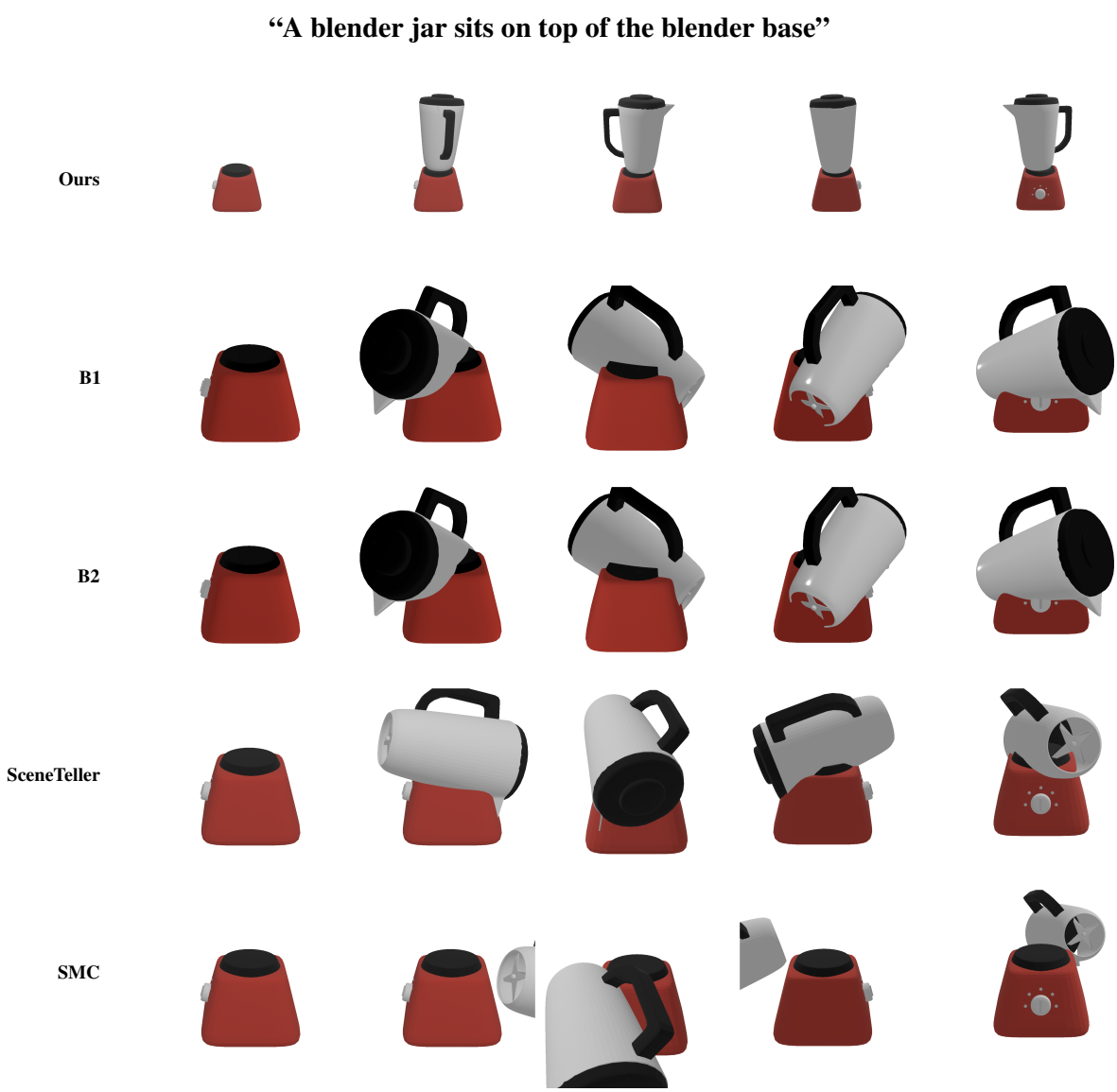


Figure 20. **Per-mesh qualitative comparison: blender jar and base.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Only our method places the jar upright and stably on the base; other baselines do put the jar on top but typically tilt it and exhibit noticeable interpenetrations between jar and base.

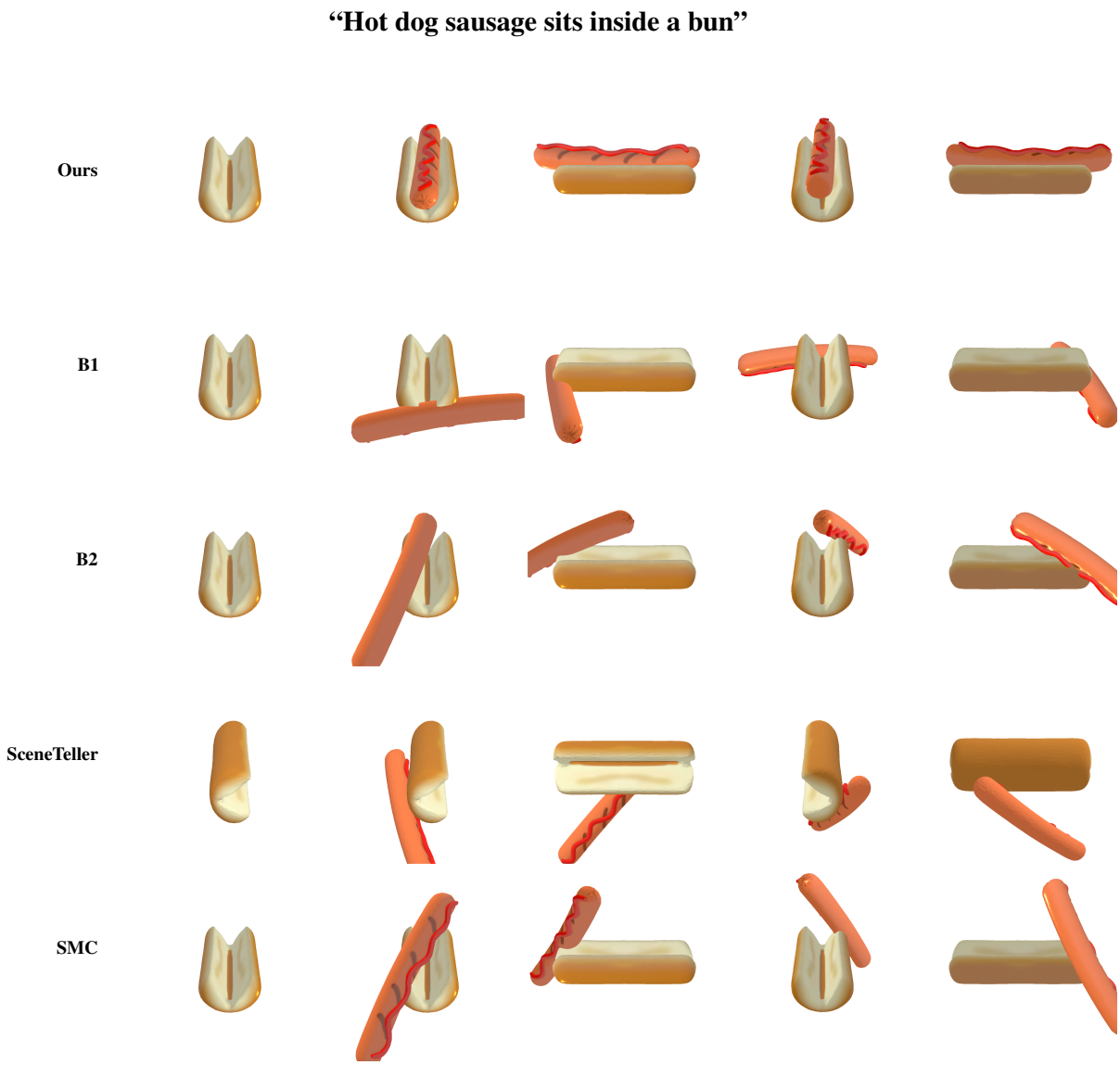


Figure 21. **Per-mesh qualitative comparison: hot dog and bun.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Only our method produces a physically plausible configuration where the sausage rests inside the bun; B2 and SMC place the hot dog above the bun, but it floats without realistic contact.

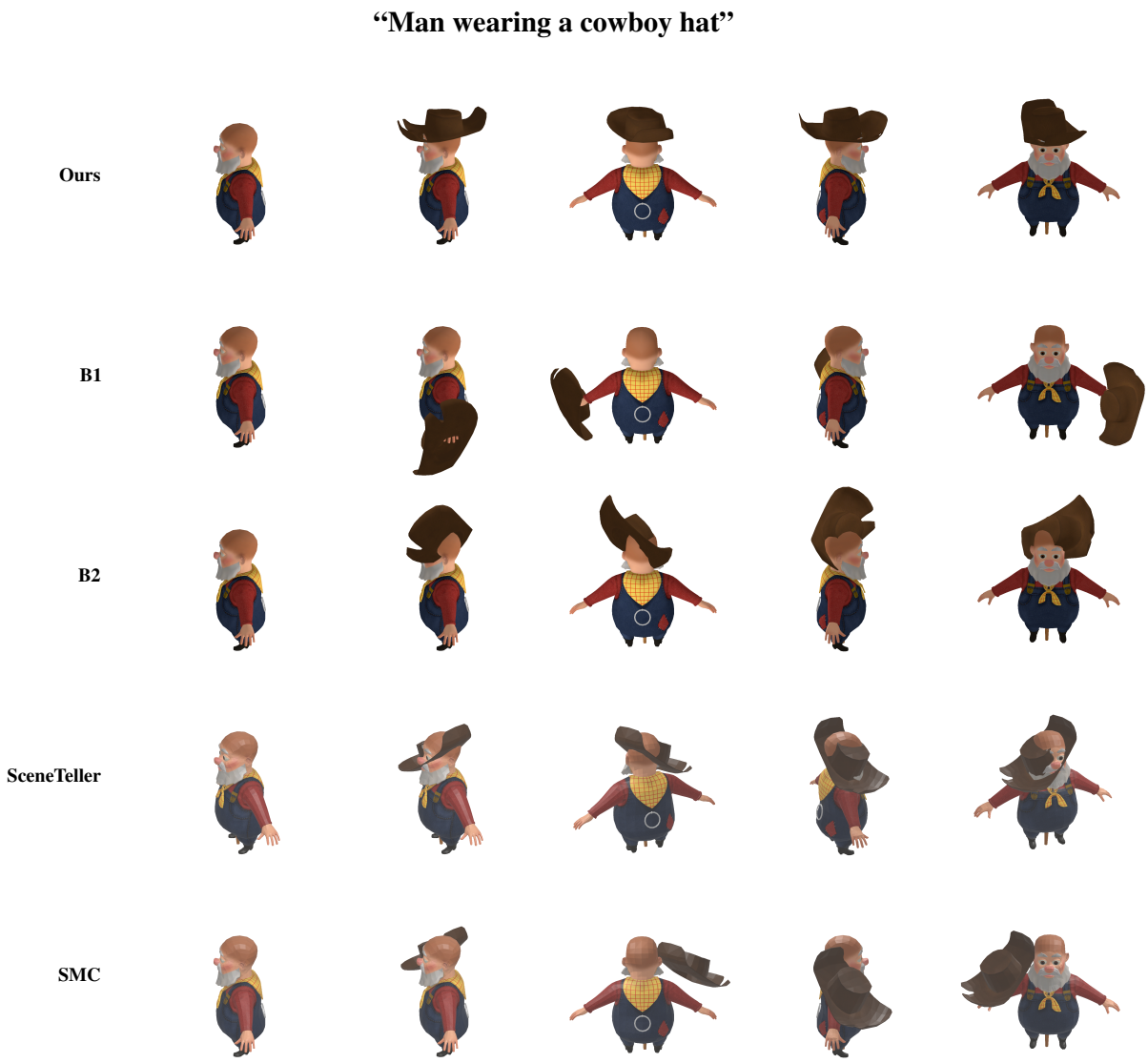


Figure 22. **Per-mesh qualitative comparison: man and cowboy hat.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Our method positions the hat stably on the head; although B2, SMC, and SceneTeller place the hat close to the man’s head, their configurations remain physically implausible, with poor contact or penetration.

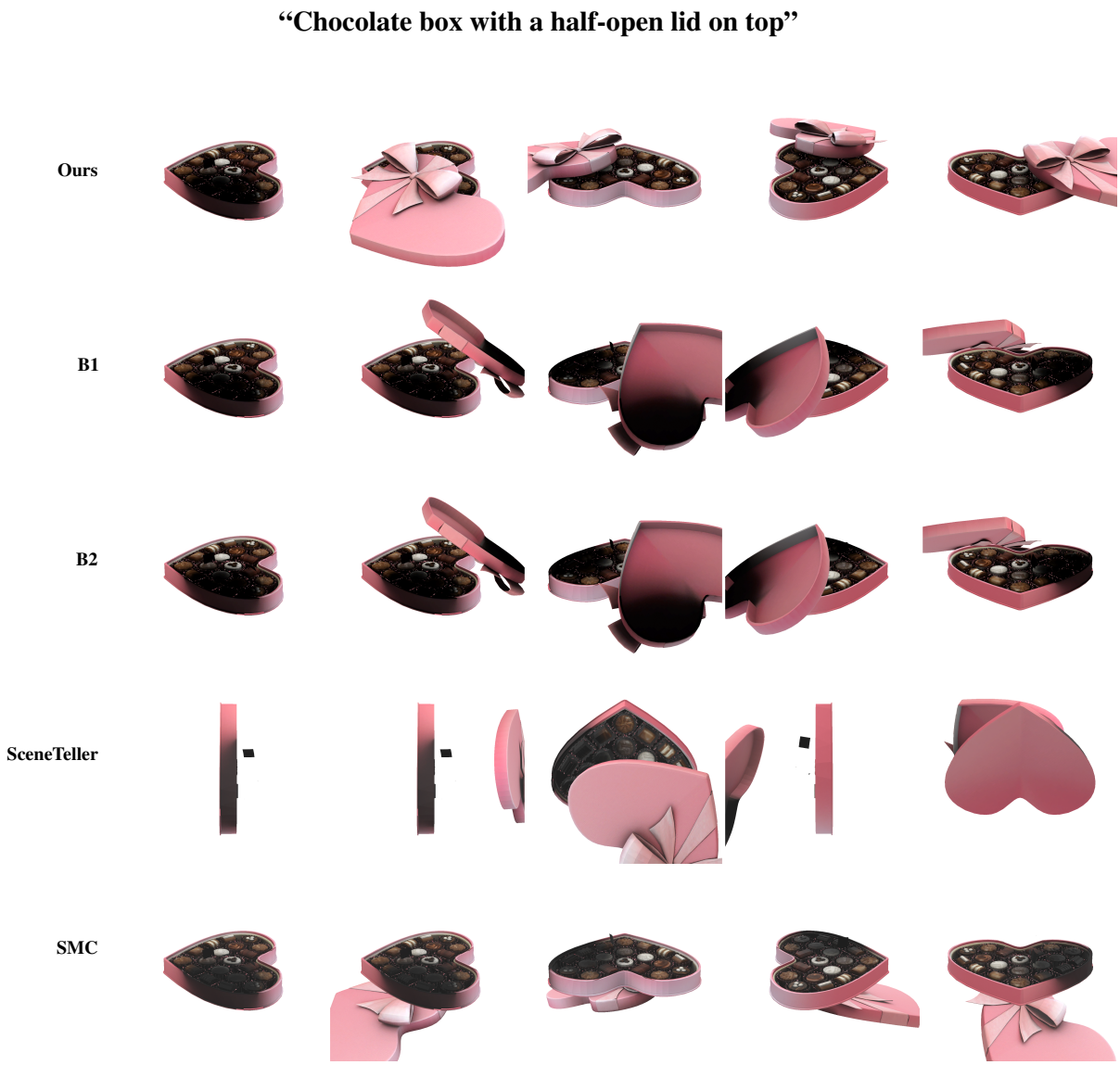


Figure 23. **Per-mesh qualitative comparison: chocolate box and lid.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Only our result matches the prompt with a half-open lid resting on the box; in B1 and B2 the lid appears upside down, in SceneTeller it floats above the box, and in SMC it is placed below the box rather than on top.

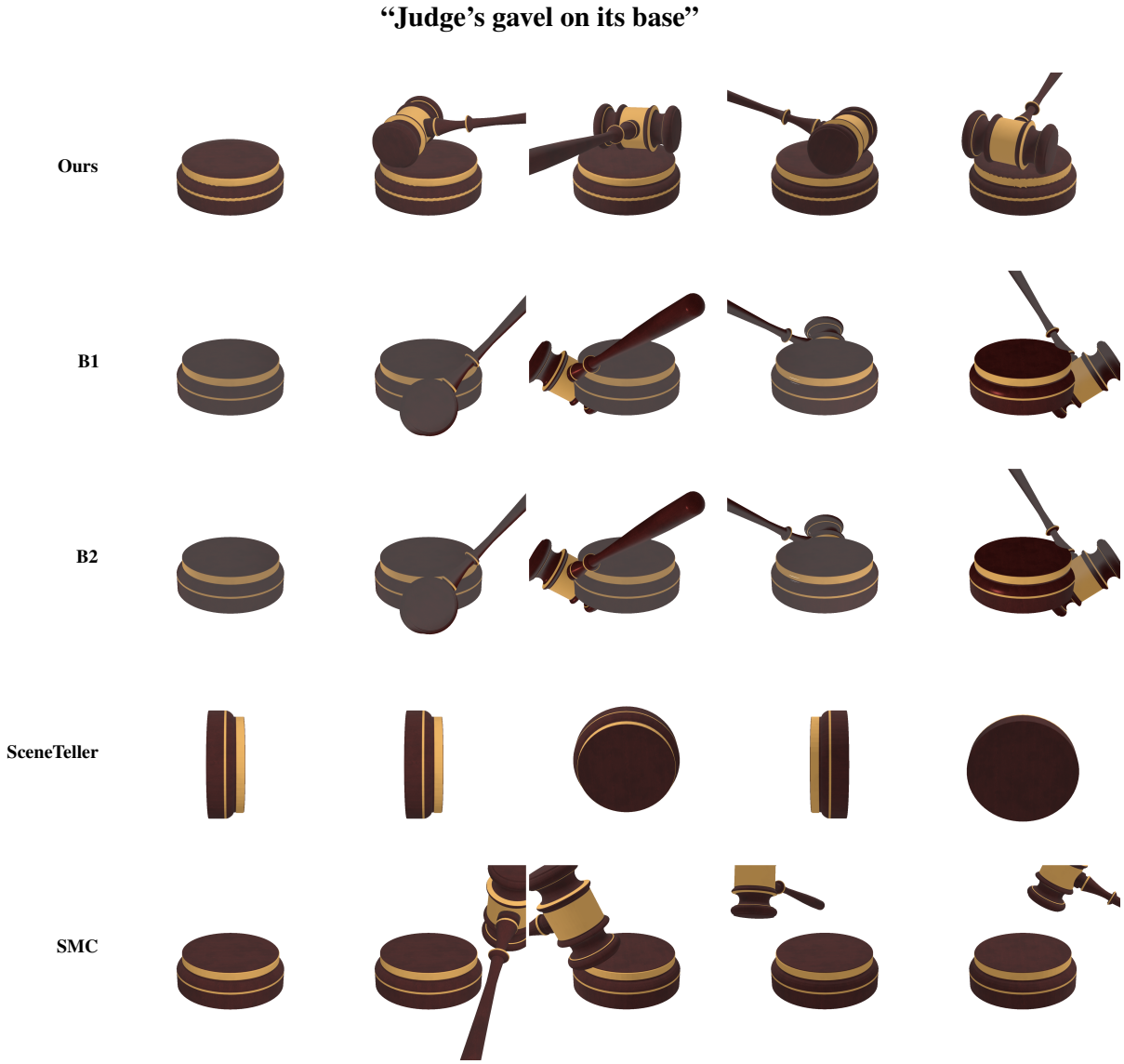


Figure 24. **Per-mesh qualitative comparison: judge’s gavel and base.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. Our method places the gavel directly and stably on its base; in B1 and B2 the gavel misses the base. Scene teller place the gavel far away from it’s base, and in SMC the configuration can look text-aligned from some views but the gavel is not actually resting above the base.

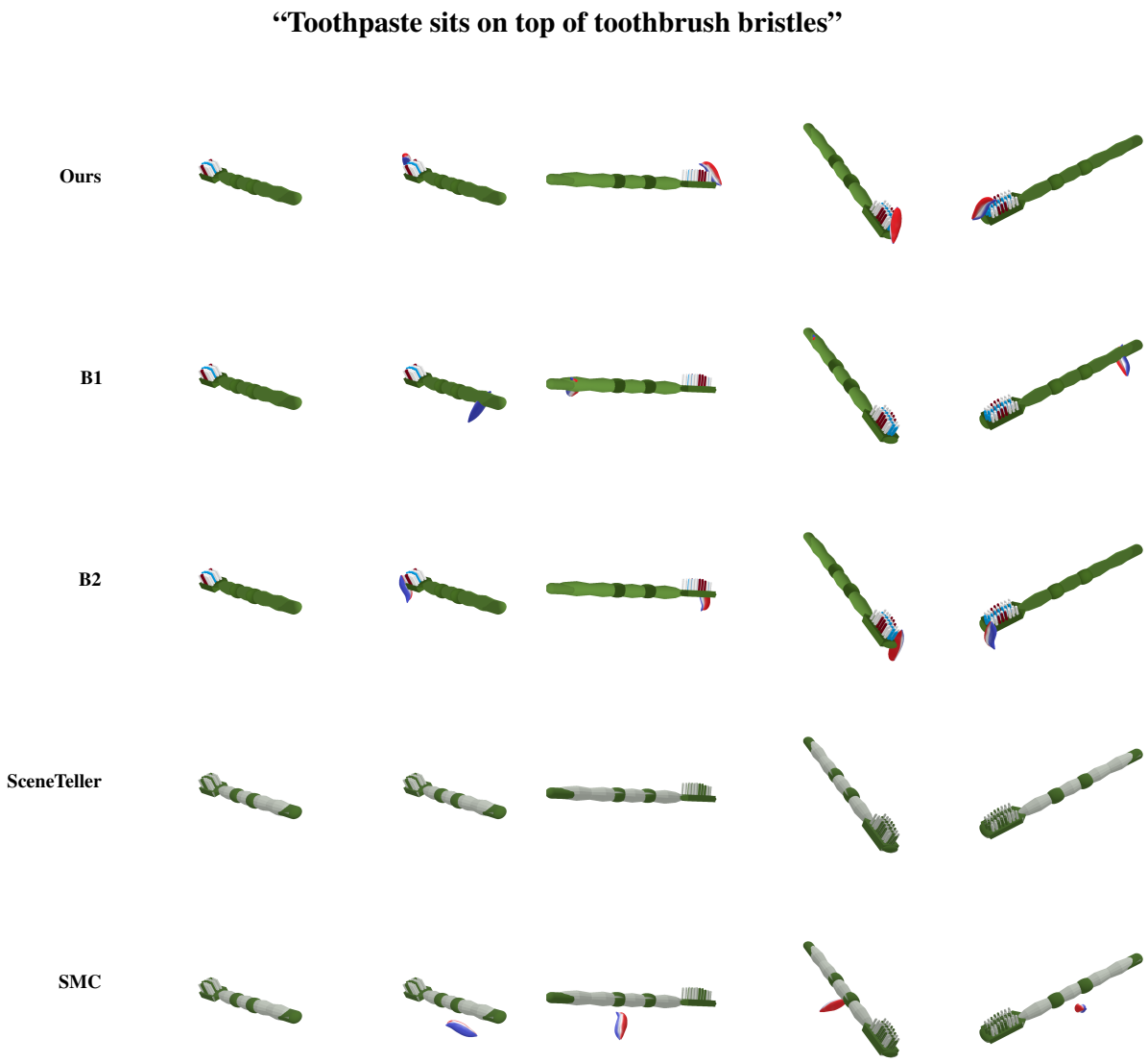


Figure 25. **Per-mesh qualitative comparison: toothpaste and toothbrush.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. While not perfect, our method yields a physically plausible placement of the toothpaste on the bristles. B2 comes close to placing it on top but still produces a floating, non-physical configuration.

Figure 26. **Per-mesh qualitative comparison: queen and crown.** Rows correspond to our method, B1, B2, SceneTeller, and SMC (top to bottom). In each row, the leftmost panel shows the target mesh alone, followed by the final aligned meshes rendered from four different viewing angles. In our result, the crown sits directly and stably on the queen’s head; in the other baselines the crown is either far from the head, or even not clearly visible in the final view.