# FAME: 3D Shape Generation via Functionality-Aware Model Evolution (Supplementary Material)

Yanran Guan, Han Liu, Kun Liu, Kangxue Yin, Ruizhen Hu, Oliver van Kaick,
Yan Zhang, Ersin Yumer, Nathan Carr, Radomir Mech, Hao Zhang

In this document, we present additional evaluation and results of our method, including details on user studies.

## S1 Additional Results

Fig. S1 shows additional results of functionality-aware model evolution for an input population of four shapes.

## S2 Examples of Beam Search

Given a hybrid shape possessing two functionalities, Fig. S2 and Fig. S3 show the complete beam search conducted on the shape and the optimal partial shapes found for two functional categories: *chair* and *desk*.

## S3 User Studies

*Functionality partial matching and multi-functionality*. To evaluate functionality partial matching and the multi-functionality measure, we designed a study consisting of 30 functional hybrids produced by our modeling tool with a varied mix of functional categories covering the initial population. For each hybrid, a participant was asked to choose, among six candidate functional categories, "all functions that you believe the object supports". We collected responses from non-expert users through Amazon Mechanical Turk, obtaining 32 responses in total.

We check the consistency between the responses and the functionalities detected by our partial matching and multi-functionality score. For each hybrid $\mathcal{H}$, each functionality label is considered as assigned to $\mathcal{H}$ by the users if at least $20\,\%$ of users selected the label for $\mathcal{H}$, which results in a set of labels assigned to $\mathcal{H}$, denoted as $\mathcal{L}_u^{\mathcal{H}}$. The set of functionality labels selected by our method based on partial matching and the multi-functionality measure is denoted as $\mathcal{L}_o^{\mathcal{H}}$. To compare these two sets of labels, we compute the agreement between $\mathcal{L}_u^{\mathcal{H}}$ and $\mathcal{L}_o^{\mathcal{H}}$ in terms of recall and precision. We define recall as the percentage of labels in $\mathcal{L}_u^{\mathcal{H}}$ that also appear in $\mathcal{L}_o^{\mathcal{H}}$, i.e., found by our method, and define precision as the percentage of labels in $\mathcal{L}_o^{\mathcal{H}}$ that appear in $\mathcal{L}_u^{\mathcal{H}}$, which indicates whether each of the labels we found is considered to be correct by most of the users.

TABLE S1: Various statistics from our functionality-aware model evolution experiments. $M$ denotes the number of 3D objects in the input set. $G$ denotes the total number of part groups obtained from the input set. $\%B$ denotes the average percentage of offsprings produced which broke symmetries from their parents. Precision and recall denote the user agreement with our functional plausibility scores computed for the input set.

| Input set | $M$ | $G$ | $\%B$ | Precision | Recall |
|---|---|---|---|---|---|
| Fig. 1 | 4 | 21 | $25\,\%$ | 0.98 | 0.97 |
| Fig. 11 left | 4 | 24 | $45\,\%$ | 0.98 | 0.89 |
| Fig. 11 right | 4 | 17 | $6.4\,\%$ | 0.80 | 0.80 |

We compute the average precision and recall for the 30 hybrids in the study, obtaining 0.92 and 0.89, respectively. These relatively high values indicate that our functionality partial matching and the multi-functionality measure can effectively detect most of the partial functionalities that the hybrids support.

A breakdown of the precision and recall for two of our modeling sessions (Fig. 1 and Fig. 11 in the paper) is given in Table S1. Note that the precision and recall for the input shown in Fig. 11 right are relatively low compared to those of the other sets. We believe that the main reason is that we denoted the main functionality of the *shelf* category using the *placement* functionality label, while users tend to assign the label *storage*. Due to this discrepancy of interpreting the meanings of the functionality labels, the sets of labels selected by our method and the users had some inconsistency.

*Functional plausibility user study*. To assess our functional plausibility measure from the viewpoint of users, we conducted a user study. The study involved 20 pairs of offsprings generated by our modeling evolution: each pair was *randomly* selected from three generations of three populations. The sampled shapes possess a range of functionalities and plausibility scores, including shapes with low plausibility scores, and a range of distances between the plausibility scores in each pair. For each pair, a participant was asked to provide a relative evaluation of the plausibility of the

Fig. S1: Starting from a heterogeneous collection of four objects (in gray) as initial population, our functionality-aware evolutionary modeling tool is able to generate a variety of offspring shapes (in yellow). Some of the offspring shapes exhibit forms of cross-category *structure breaking*.

shapes by choosing which model he/she thinks is more functionally plausible. This was described in the question as “the model that better supports its intended functionality in comparison to the other model”. For each question, the users could select among three options: “the shape on the left is more plausible”, “the shape on the right is more plausible”, or “both shapes are equally plausible”.

In the end, $32$ Turkers provided $640$ responses. For each pair, if either of the first two options is selected by more than $50\,\%$ of the users, which implies that the majority of the users agree that one shape is more plausible than the other, then we evaluate the agreement between the user responses and our plausibility score, by checking whether the hybrid considered more plausible by the users has a higher plausibility score than the other shape. Otherwise, we consider that there is no consistent agreement among users and the two shapes are considered to be equally plausible. In this case, we verify whether our plausibility score agrees with the user responses by checking if the difference between the plausibility scores of the two shapes is smaller than a threshold of $0.05$. By adding up the agreement between scores and user responses for these two cases, we find that $13$ ($65\,\%$) of the responses among the $20$ pairs are consistent with our functional plausibility score.

To study why the consistency is relatively low, we computed the information entropy of all the votes for each pair, obtaining an average entropy of $1.31$ for all the pairs. Note that, if all the users consistently agreed in one of the three options, then the entropy would be zero. On the other hand, if all the three options were equally selected by the users, implying $1/3$ of the votes for each option, then the entropy would be maximal, with a value of $\log_2 3 = 1.58$. We see that the entropy of the responses obtained from the users is close to the entropy of random decision, which would indicate that evaluating the functionality plausibility of objects is a difficult task for humans, especially for shapes with multiple functionalities.

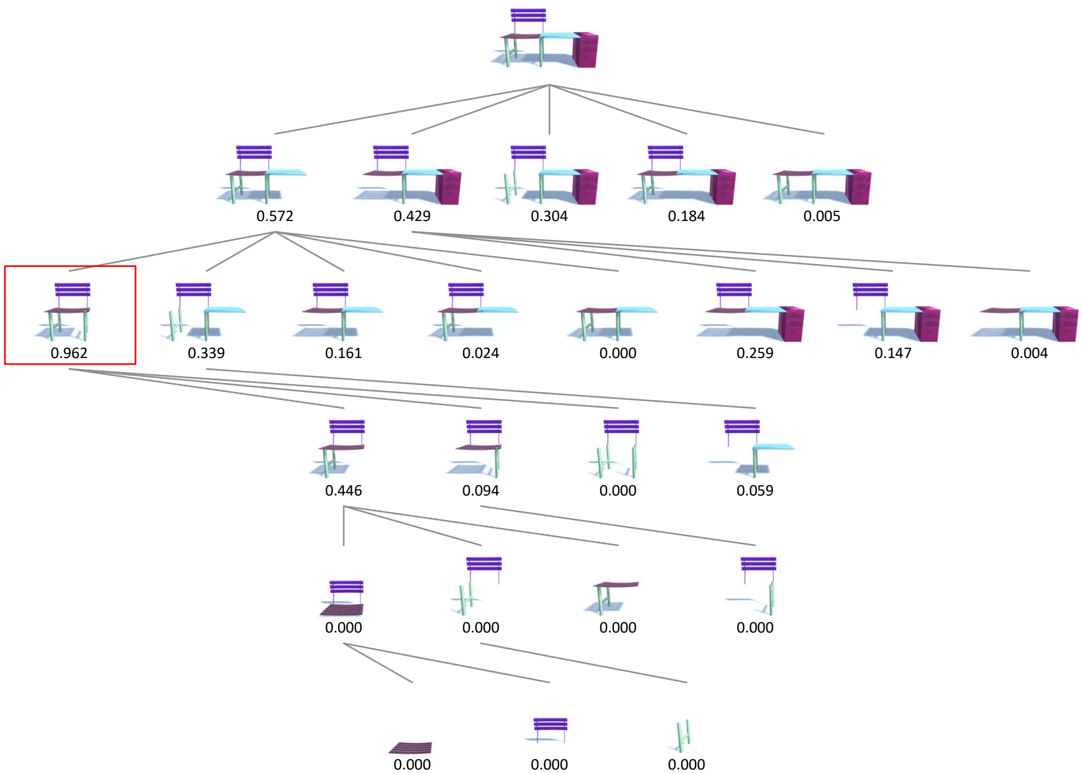


Fig. S2: Beam search for functionality partial matching, where we search for the subset of parts that provides the highest score for the *chair* category.

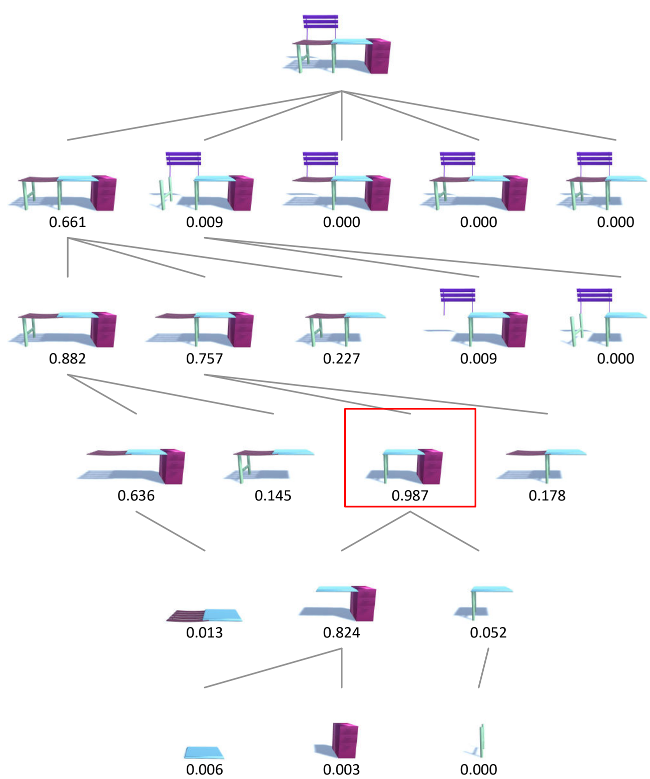


Fig. S3: Beam search for functionality partial matching, where we search for the subset of parts that provides the highest score for the *desk* category.