Supplementary Material: Property-Informed Diffusion-Based Text-to-Microstructure Generation

Bingxuan Dai[1 †], Hongsong Wang[2,3 †*], Jie Gui[1,4,5 *]
[1]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4]Purple Mountain Laboratories, Nanjing 210000, China
[5]Engineering Research Center of Blockchain Application, Supervision And Management (Southeast University), Ministry of Education, China
{220245799, hongsongwang, guijie}@seu.edu.cn

A. Test-Time Reward-Guided Alignment

In the reward model, we employ a pre-trained CLIP-style dual encoder and discriminator to jointly evaluate the semantic alignment and structural plausibility of the generated microstructures, which is illustrated in Figure 1. The discriminator is trained to distinguish high-quality text-aligned structures from generated samples. The discriminator's local structure score is denoted as s_p, the discriminator's global structure score is denoted as s_g, and the overall score s_f is obtained by averaging the patch-level and global probabilities. The weights of CLIP and the discriminator are 1.0 and 0.4, respectively.

Detailed process of test-time reward-guided alignment is shown in Algorithm 1. The initial sample undergoes $R = 5$ iterative optimization steps. In each iteration, $E = 4$ local patches of size $32 \times 32 \times 32$ are randomly selected as candidate editing regions. For each patch, $K = 8$ candidate structures are generated. These candidates are evaluated using a reward model that weights the semantic similarity score from CLIP and the structural realism score from the discriminator. For each local region, the candidate with the highest reward score replaces the original patch. After each round of optimization, the current best reward is normalized using a softmax function, and the structure from the pool of best resampled samples is used as the starting point for the next round. This iterative process continues until a predetermined number of rounds is reached.

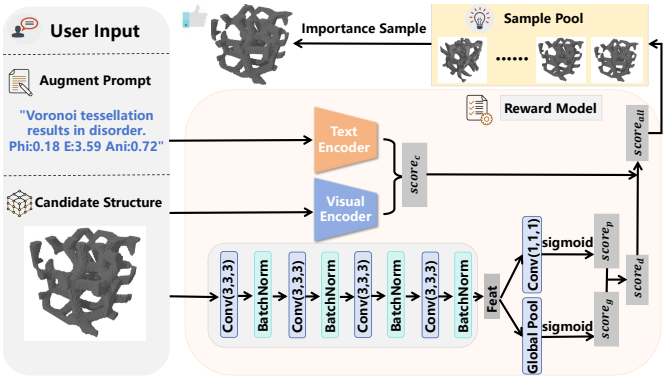

Figure 1. **Overview of Test-Time Reward-Guided Alignment:** Given an initial material structure and a target textual description, the proposed module performs localized optimization of random regions guided by metric scores derived from two reward models: a contrastive reward and a discriminative reward, thereby progressively refining the 3D material structure toward the target.

B. Evaluation Metrics

To evaluate the performance of our generated microstructure, we conduct a quantitative analysis using four complementary metrics: classification accuracy, Fréchet Inception Distance (FID), CLIP score, and Chamfer Distance (CD). These metrics respectively assess the semantic consistency, distributional fidelity, text-to-structure alignment, and geometric accuracy of the generated microstructures. Furthermore, we used the R-squared coefficient to evaluate the correlation between the predicted properties of the generated structure and the predicted properties of the real structure.

*Corresponding Authors. †Equal Contribution

Algorithm 1 Test-Time Reward-Guided Local Sampling

Input: Diffusion model $\mathcal{G}$, prompt T, batch size B, diffusion steps S, guidance scaling factor w, edits per round E, candidates per edit K, total rounds N, temperature λ
Output: Optimized structures X^*
Initialize X from the diffusion model $\mathcal{G}(T, B, S, w)$
Set $X^* \leftarrow X$, and compute initial rewards R^*
for $n = 1$ to N **do**
  Sample E patch locations
  for $e = 1$ to E **do**
    for $k = 1$ to K **do**
      Local edit X with half-step sampling $\hat{X}^{(k)}$
      Replace patch e with a new sample
    end for
    Evaluate the combined rewards of K candidates
    Update X with the highest reward
  end for
  Compute current rewards R
  if Reward enhancement **then**
    Update X^* and R^*
  end if
  Compute probabilities $p = \mathrm{softmax}(\lambda R^*)$
  Sample indices $s \sim \mathrm{Multinomial}(p, B)$
  Set $X \leftarrow X^*[s]$
end for
return X^*

B.1. Classification Accuracy

It can evaluate the ability of the model to generate microstructures that are recognizable and distinguishable across predefined categories. To perform this evaluation, a 3D convolutional neural network classifier is trained using 2,000 labeled microstructures distributed over 20 semantic classes. The high accuracy achieved by the classifier on the generated samples indicates strong semantic alignment between the structures and the textual prompt.

B.2. FID Score

To evaluate the generation quality of 3D microstructure, we adopt the Fréchet Inception Distance (FID) metric by extracting feature embeddings from the penultimate layer of a pre-trained 3D classifier. We compute the FID score between 2,000 real and 2,000 generated microstructures using the standard formulation:

$$FID = ||\mu_x - \mu_y||^2 + Tr(C_x + C_y - 2(C_x C_y)^{1/2}) \quad (1)$$

where x and y represent the feature vectors of real and generated structures, respectively, μ_x and μ_y are the mean vectors, C_x and C_y are the covariance matrices, and Tr denotes the matrix trace. A lower FID indicates that the distribution of generated structures is closer to that of real samples, suggesting higher generation fidelity.

B.3. CLIP Score

It can measure the degree of semantic alignment between the input text and the corresponding microstructure generated. A pretrained dual encoder from the first stage of training is employed to embed both the text description and the generated microstructure into a shared representation space. Semantic alignment is subsequently assessed by computing the cosine similarity between their corresponding embeddings. A high similarity score indicates that the model effectively captures and preserves the semantic content of the input prompt in the generated output.

B.4. CD Score

It is computed between point clouds sampled from the generated and ground-truth material structures, serving as a measure of geometric similarity. To compute the CD score, the material structures, both generated and ground-truth, are first converted into surface point clouds, typically by sampling the coordinates of boundary voxels or using isosurface extraction methods. Let $P = \{p_i\}_{i=1}^N$ and $Q = \{q_j\}_{j=1}^M$ represent the two point clouds sampled from the generated and ground-truth structures, respectively. The CD score is written as:

$$CD = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|q-p\|_2^2 \quad (2)$$

which ensures that each point in one set is close to at least one point in the other. The first term evaluates how much of the generated structure is relevant to the target geometry, while the second term captures how well the generated structure covers the target. The lower CD score corresponds to higher shape fidelity and more accurate reconstructions.

B.5. R-Squared Coefficient

It can evaluate the ability of the model to generate microstructures that closely match the target physical properties. For this evaluation, we train a solver with the same convolutional architecture as the classifier using 2000 properties-labeled microstructures. A higher R-squared coefficient, approaching one, indicates stronger physical consistency in the generated structures.

C. Implementation Details

During the pre-training phase, PropDiff-TMG trains a property-informed dual encoder using a contrastive text-structure pre-training approach to align text-structure embeddings. The dataset contains $N = 2000$ microstructures, with 80% used for training and 20% for validation. Each text description is accompanied by the physical properties

of its corresponding microstructure, including volume fraction, effective Young's modulus, and isotropy. To improve generalization, the physical property is randomly masked during training. A pre-trained 3D-ResNet model [2] serves as the structure encoder, and a pre-trained BERT model [1] serves as the text encoder. The two encoders are jointly trained to optimize the similarity matrices of structure and text embeddings. The model trains within a self-conditional diffusion framework, processing input microstructures at a resolution of $64 \times 64 \times 64$ voxels. Training is performed using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 64, and a temperature parameter of 0.1. Attention mechanisms operate at spatial resolutions of 4 and 8, utilizing 4 attention heads and a dropout rate of 0.1. Training incorporates an exponential moving average (EMA) with a decay rate of 0.999 to stabilize parameter updates. Training runs for 2000 epochs on a single NVIDIA GeForce RTX 4090 GPU. The inference time is approximately 5 seconds per sample with 100 diffusion steps on a single RTX 4090 GPU, while the optimized testing-time inference takes 110.4 seconds.

D. Experiment

D.1. Dataset Construction

GenText-Microstructure is a text-structure pair dataset based on automatically generated microstructure physical descriptions from 3D voxels [3] as prompts. Over 14,000 metamaterial structures spanning a broad spectrum of modulus and Poisson's ratio are utilized for training, while another 2,000 randomly generated structures are used for evaluation. For each binary voxel, we extract geometric features such as connectivity, Euler number, surface area to volume ratio, symmetry, and internal porosity. In addition, the corresponding stiffness tensor provides mechanical properties, including C_{11}, C_{12}, C_{44}, and the derived bulk modulus V. All data are normalized using a pre-trained scaler. Subsequently, GPT is employed to generate descriptive text conditioned on the properties and features, and a rule-based text generator converts the quantitative properties into a coherent, scientifically styled natural-language description, covering volumetric stiffness, axial stiffness, Poisson's ratio, shear strength, and structural topology.

D.2. Visualizations and Simulations

Qualitative Analysis on Geometries 2000: Figure 2 shows the visualization of qualitative results generated based on properties-driven text. The introduction of property constraints makes the generated results more targeted while still meeting the target physical properties, but this also reduces structural diversity.

Qualitative Analysis on GenText-Microstructure: To evaluate the model capacity to interpret and translate structured textual descriptions into diverse material geometries, we conducted qualitative visualization experiments using the GenText-Microstructure. As shown in Figure 3, the visualization results demonstrate that the model effectively captures the underlying structural semantics while exhibiting diverse results. As shown in Figure 4, the generated structures after adding properties maintain the overall form while exhibiting local diversity. This indicates that the model not only learns deterministic text-structure mappings but can also generalize to generate different structures under the same text semantics.

Visualization of Simulation: Simulation results are presented in the videos included in the supplementary materials. The compression process of the resulting metamaterial structure is visually recorded via video contour plots extracted from a series of nonlinear finite element simulations. These contour plots illustrate the gradual evolution of deformation as the axial compressive displacement is incrementally applied along the z-axis, starting from 0 to 0.3. The video begins with the undeformed configuration and proceeds through successive loading stages. Throughout the simulation, the displacement field distribution is colormapped to emphasize regions of high deformation gradients. The progression demonstrates that the generated structure consistently exhibits auxetic behavior, as indicated by lateral compression in the x- and y-directions under axial compression. Compared to the reference metamaterial, the generated structure maintains a pronounced negative Poisson's ratio effect across a wide range of compressive strains. Furthermore, the contour plots reveal localized stress concentrations and deformation distortions. Overall, the results confirm that the generated structures maintain the desired mechanical behavior.

D.3. Visualization under Diverse Prompts

To evaluate model generalization and text–structure alignment beyond the training distribution, we conduct a qualitative visualization experiment using newly generated and diverse text prompts created by GPT. The Geometries 2000 [4] includes four types of materials: metals and alloys, polymers and composites, ceramics, building materials, and metamaterials. For each material category, we select a representative target structure and textual descriptions generated by GPT describing its morphology and spatial distribution. The model then generates 3D structures based on each textual prompt. As shown in Figure 5, the generated samples demonstrate that the model can accurately capture structural semantics from various textual descriptions. This confirms that the model can generalize beyond memorized training samples and align textual semantics with structure generation in an interpretable manner.

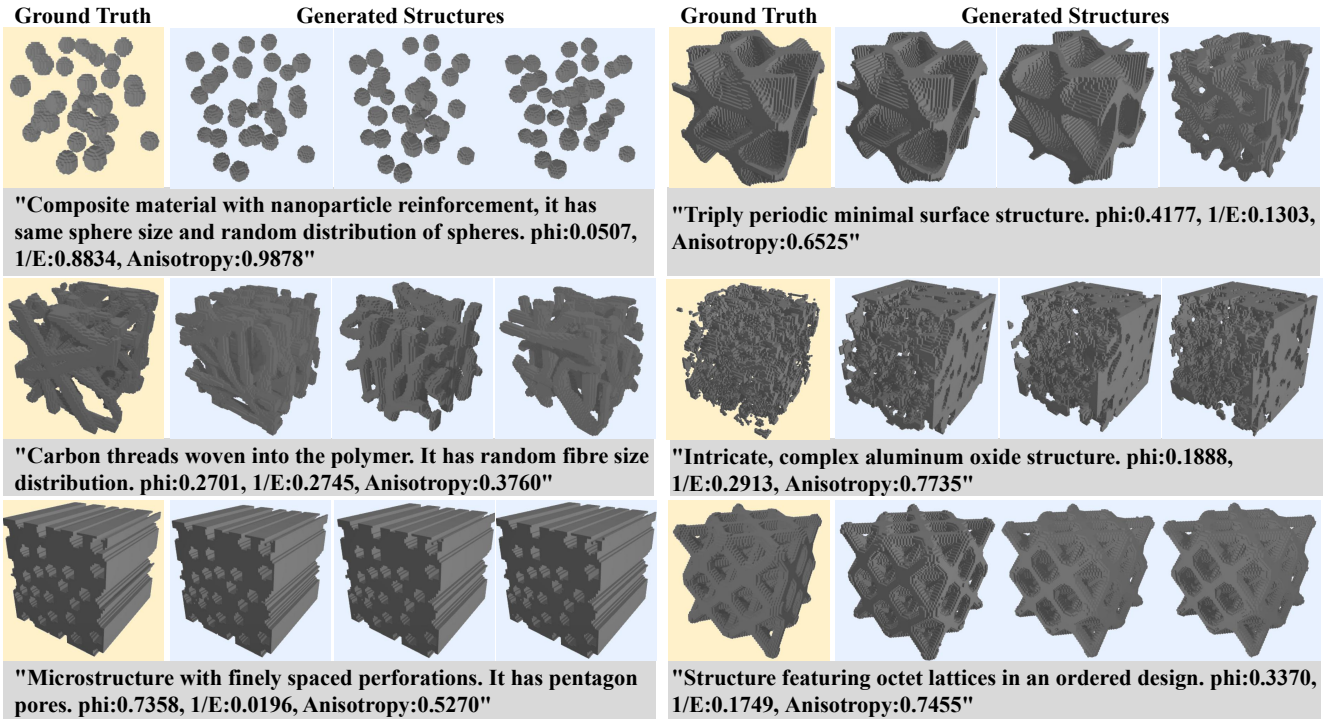

Figure 2. **More qualitative visualizations:** Voxel-based microstructures generated by our model using textual prompts with properties.

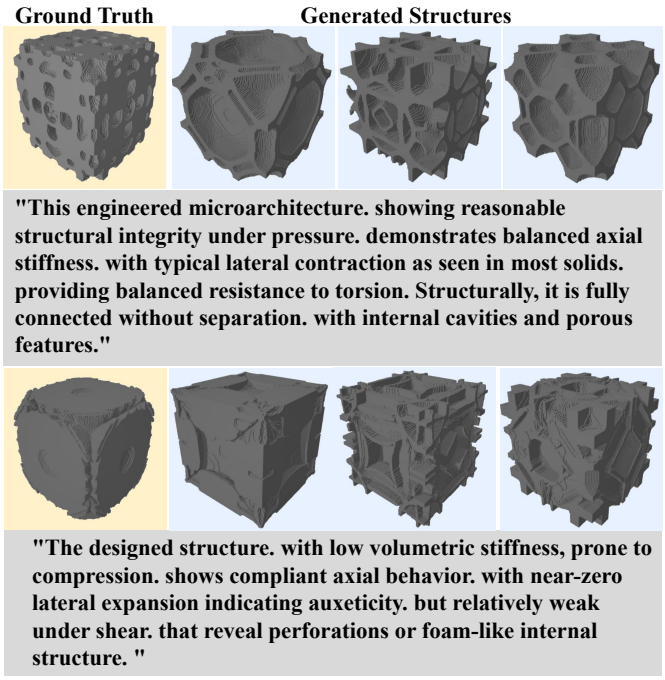

Figure 3. **Qualitative visualizations of GenText-Microstruture without physical properties:** Voxel-based mechanical metamaterial structures are generated in response to text prompts produced by a rule-based text generator.

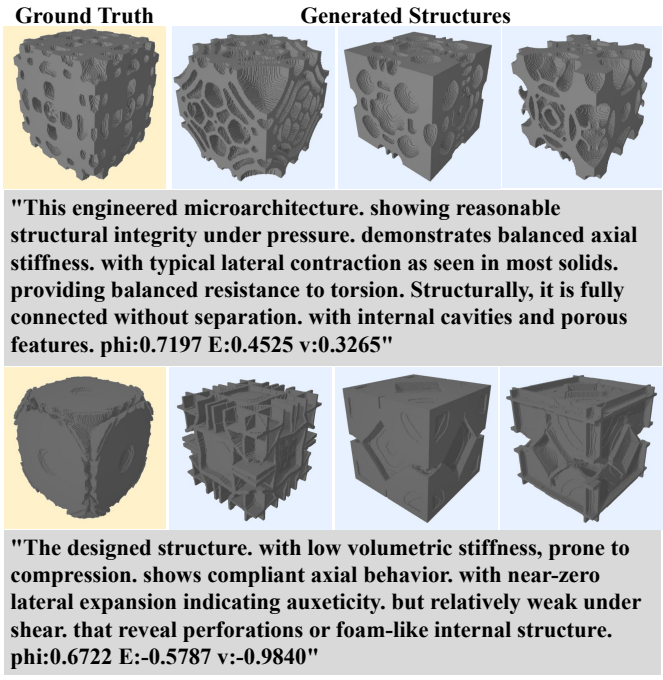

Figure 4. **Qualitative visualizations of GenText-Microstruture with physical properties:** Voxel-based mechanical metamaterial structures are generated based on attributed text prompts generated by a rule-based text generator.

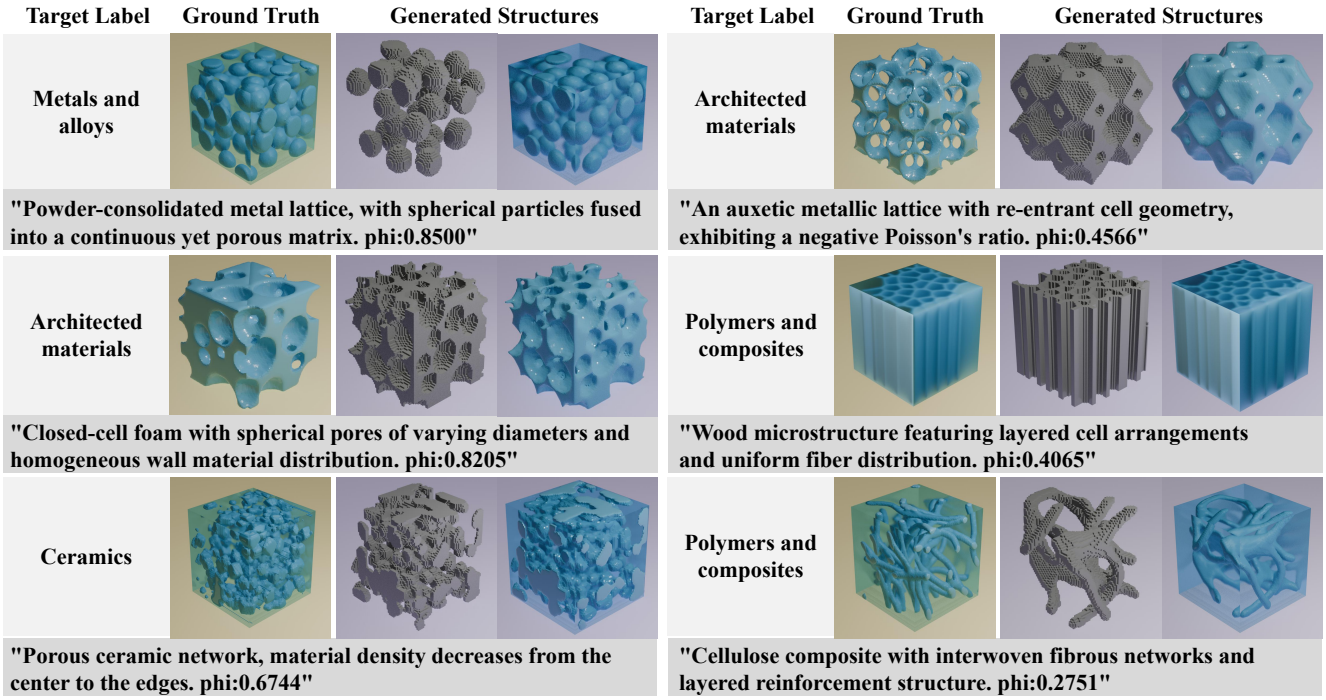

Figure 5. **Qualitative visualizations of GPT-generated textual descriptions:** 3D structure generation was conducted based on GPT-generated textual descriptions across different material categories. The figure presents representative target structures, the corresponding GPT-generated text prompts, and the resulting generated 3D structures.

D.4. Ablation Experiment

To comprehensively evaluate the contribution and generalization ability of different components in our method, we conduct ablation studies on different modules on GenText-Microstruture, as shown in Table 1.

Effectiveness of Property-Informed Stochastic Conditioning: To verify the effectiveness of property-informed stochastic conditioning, we conduct additional ablation experiments on our constructed text-structure dataset. As shown in Table 1, the results show that adding physical properties further improves the model across all metrics, indicating higher consistency in semantics among the generated structures. Furthermore, this module effectively reduces ambiguity in structure generation, enabling the model to more accurately align expected physical features while maintaining diversity.

Effectiveness of Test-Time Reward-Guided Alignment: To validate the effectiveness and generalization ability of the test-time reward-guided alignment module, we conduct additional ablation experiments on our constructed text-structure dataset. The text prompts on this dataset are generated using regularized templates, systematically describing different structural features, and differed from the original dataset in text-structure distribution. As shown in Table 1, the results show that the proposed reward-guided alignment still significantly improves performance across multiple metrics, further confirming that this module can guide the model to generate structures that are more functionally and physically consistent without additional training.

Table 1. **Ablation studies:** Contrastive Align and Reward-Guided Align denote Contrastive Text–Structure Alignment and Test-Time Reward-Guided Alignment, respectively. In evaluating the FID metric, we select the visual encoder from Contrastive Text-Structure Alignment as the feature extractor.

	Method	FID ↓	CLIP ↑	CD ↓
①	Ours	**47.74**	**0.6463**	**0.0442**
②	w/o Property condition	52.94	0.5210	0.0482
③	w/o Reward-Guided Align	49.02	0.5164	0.0468
⑤	w/o Discriminator	51.89	0.6396	**0.0404**
⑥	w/o Normalization	51.94	0.6396	0.0454

Furthermore, additional ablation experiments are conducted on each component of the module to validate the effectiveness of the discriminator and the normalized combined reward function. As shown in Table 1, the results demonstrate that incorporating the discriminator slightly increases the CD score but improves FID and CLIP, indicating enhanced diversity and semantic fidelity instead of target memorization. Meanwhile, the normalized combined reward function balances the contributions of the two reward

metrics, avoiding a single factor dominating the optimization process, thereby improving all metrics. This further demonstrates the effectiveness of these two modules.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 3

[2] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[3] Yanyan Yang, Lili Wang, Xiaoya Zhai, Kai Chen, Wenming Wu, Yunkai Zhao, Ligang Liu, and Xiao-Ming Fu. Guided diffusion for fast inverse design of density-based mechanical metamaterials. *arXiv preprint arXiv:2401.13570*, 2024. 3

[4] Xiaoyang Zheng, Ikumu Watanabe, Jamie Paik, Jingjing Li, Xiaofeng Guo, and Masanobu Naito. Text-to-microstructure generation using generative deep learning. *Small*, 20(37): 2402685, 2024. 3