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`

Abstract

Designing 3D metamaterial microstructures that meet the intended functions remains a major challenge, as it typically requires domain expertise, iterative simulations, and extensive manual tuning. Existing work on inverse design that automatically generates microstructures based on desired target properties often suffers from limited design diversity and faces challenges in ensuring the physical feasibility of the generated structures. To address this issue, a property-informed diffusion-based network is proposed that enables the generation of 3D microstructures directly from textual descriptions. Unlike traditional property conditioning methods, our approach leverages rich guidance in terms of semantics and physical properties in the text input to support diverse structure synthesis. To enforce consistency between the generated structures and the target textual prompts, a dual alignment strategy is adopted, including contrastive text-structure alignment and test-time reward-guided alignment. Experimental results show that the model is capable of generating semantically meaningful and physically plausible structures across a wide range of material categories. Our approach has good potential for interactive microstructure design and opens up new directions for combining language-based interfaces with inverse material discovery. Code is available at: `https://github.com/hongsong-wang/PropDiff-TMG`.

1. Introduction

Emerging materials [6, 12, 18, 52], characterized by distinctive microstructures, enhanced performance, and novel functionalities, have garnered increasing attention due to the rapid advancement of scientific research and engineering innovation. Their integration into high-technology sectors such as aerospace [31], environmental protection [32], and biomedical engineering [43] has substantially expanded both the functionality and application potential of materials. The properties of these materials are predominantly governed by their internal microstructures rather than their constituent substances. Consequently, the design and engineering of microstructures have assumed a central role in the development of novel materials, enabling the optimization of performance and the customized realization of specific functionalities.

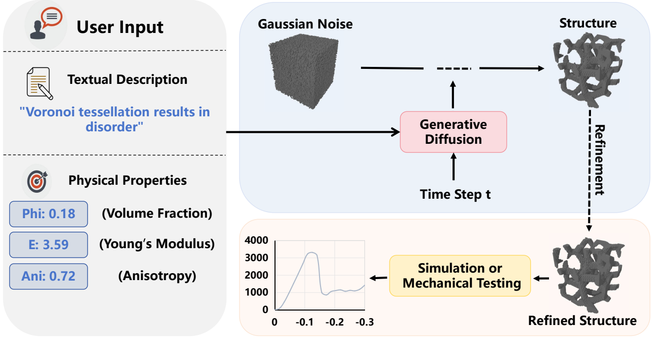


Figure 1. **Overview of our approach:** A diffusion-based framework can generate metamaterial microstructures according to the textual description and physical properties.

Inverse design has emerged as a powerful paradigm in materials science, shifting the traditional trial-and-error process toward a goal-driven approach [17, 18, 49, 60]. In mechanical metamaterials research, phase-field modeling [56] and topology optimization [39, 51] have been used

---

[*]Corresponding Authors. [†]Equal Contribution

to inversely design architectures with extraordinary characteristics such as extreme stiffness or auxetic behavior [20]. Meanwhile, data-driven and machine learning techniques have accelerated inverse design by enabling rapid exploration of vast microstructural design spaces [18, 26, 29].

Given the specific physical properties of the microstructure, experienced material scientists can inversely design the corresponding microstructures using computational approaches such as phase-field modeling [27], mathematical modeling [59], and topology optimization [62]. Although these traditional methods are highly interpretable and follow physical principles, they often rely on extensive domain expertise, hand-crafted design spaces, and computationally expensive iterative solvers. Although recent advances in deep learning have made data-driven material structure design possible, most existing methods [9, 38, 60] still rely on expert-defined conditions or parameter controls. In contrast, the ability to generate metamaterial microstructures directly from textual descriptions remains underexplored, limiting the accessibility of such models.

To address these problems, we propose a property-informed diffusion-based text-to-microstructure generation (PropDiff-TMG), a fully automated, robust, and generalizable framework for the generation of metamaterial microstructures. As shown in Figure 1, our method accepts text as input and generates high-quality microstructures.

To achieve stable and targeted microstructure generation based on structural text feature descriptions, we adopt a self-conditional diffusion framework to gradually guide the generation results to approach the semantic target during the denoising process. On this basis, we further introduce the random injection of physical properties to modulate the generation process to improve the consistency and controllability of the structure in terms of mechanical properties. Specifically, we adopt the feature-wise linear modulation (FiLM) [34] to jointly encode physical and semantic information into a dynamic adjustment signal at the feature level, thereby achieving finer-grained generation control. After training, the model can generate diverse and physically consistent 3D microstructures based on material properties and structural text descriptions.

To enhance the semantic and physical consistency of the generated structures, a dual alignment optimization strategy is proposed. During training, a contrastive text-structure alignment is introduced to align the text encoder with the visual encoder through contrastive learning to alleviate the semantic shift of the general language model [3, 36] in the material domain. During testing, a reward-guided alignment strategy is employed to enhance the synergy between semantics and physics by constructing a composite reward function that combines CLIP similarity and discriminator authenticity, and performing normalization. To further improve local fidelity, we introduce a multi-round local editing mechanism, where candidate regions are sampled and refined based on reward signals in each iteration, making high-scoring regions easier to retain and enhance, thereby optimizing the physical rationality of the structure while maintaining global consistency.

Together, these components enable our framework to generate high-quality, diverse, and physically meaningful 3D microstructures from natural language descriptions. To train and evaluate the model, we conduct experiments on a dataset of paired text descriptions and voxelized microstructures comprising multiple types of materials. Both quantitative evaluations and simulation experiments show that our approach achieves promising results in terms of both semantic alignment and physical plausibility. In summary, our contributions are as follows:

- **Innovative inverse design paradigm:** We introduce a self-conditional diffusion framework for generating metamaterial microstructures from textual descriptions and physical properties.
- **Effective cross-modal alignment mechanism:** We propose a dual alignment mechanism integrating cross-modal contrastive training and reward-guided optimization to enforce semantic and physical consistency.

2. Related Work

Inverse Design of Microstructure: Traditional methods [15, 41, 48, 55] rely on established physical principles and empirical trial-and-error processes to design 3D structures, but often suffer from long design cycles, high computational costs, and considerable algorithmic complexity [13, 19]. To address these limitations, data-driven microstructure generation [25, 28, 42], which is based on generative artificial intelligence, has emerged. For example, inverse design of metamaterials based on diffusion models [2, 53] enables end-to-end generation conditioned on mechanical properties. However, these approaches typically depend on domain-specific priors and require expert-driven condition design.

While substantial progress has been made in text-guided 3D object generation [45, 50, 57], the design of metamaterial microstructures poses greater challenges due to their intricate physical, mechanical, and functional characteristics. Current works [4, 23, 54] typically employ textual conditions to generate 2D representations of material structures, which are subsequently transformed into 3D structures via post-processing, rather than an end-to-end generative framework. Txt2Microstruct-Net [61] leverages a variational autoencoder (VAE) to generate 3D voxels and a CLIP-based module to align textual prompt with 3D voxels in the latent space. However, aligning textual prompts with 3D voxels using simple multi-layer perceptrons is challenging, and this approach requires multi-stage training. ChatMetamaterials [18] enables efficient exploration of mechanical metamaterials by introducing a large language model–driven design engine.

Diffusion-Based Material Generation: Recently, diffusion models have made breakthrough progress in multiple generation tasks [1, 16, 47, 58], significantly improving the quality and controllability of material generation. The condition-guided joint diffusion framework can effectively generate materials with complex structures and high fidelity. TGDMat [10] is proposed as a text-guided joint diffusion model to achieve synchronous generation of atomic types, coordinates, and lattice structures in periodic materials. On the other hand, in response to the problems of multi-view consistency and illumination changes, MaterialMVP [21] effectively generates multi-view physical rendering textures with invariant illumination through the attention mechanism and consistency regularization. Material Anything [24] generates physically based materials on any 3D object based on an end-to-end diffusion network. Diffusion-based methods provide an efficient and versatile framework for material generation, enabling integration of multimodal conditions like text, geometry, and physical properties beyond traditional approaches. Our work focuses on diffusion-based microstructure generation, aiming to optimize material performance through structural design.

Physically-Based Material Generation: Enforcing physical constraints in 3D generation ensures material properties follow real-world principles. 3DTopia-XL [8] generates high-resolution 3D geometry and Physically-Based Rendering (PBR) materials via diffusion for property optimization. Diffusion Renderer [30] ensures consistent lighting by jointly estimating material and illumination. MatFuse [46] and Alchemist [40] enable precise control of material properties under varying viewpoints and lighting. RichDreamer [35] enhances geometric detail and lighting consistency by jointly generating normals and depth maps. MaterialMVP [21] and SViM3D [14] improve multi-view consistency and lighting invariance. Inspired by recent material generation research, we incorporate physical property constraints into text prompts to ensure structural integrity and functionality.

3. Approach

PropDiff-TMG is a unified diffusion-based framework for the inverse design of 3D metamaterial microstructures conditioned on semantic text and physical properties, as shown in Figure 2. The entire design process is divided into three key stages to ensure controllability, generalization, and fidelity. First, we adopt a self-conditional diffusion model to generate microstructures from semantic descriptions. Second, we propose a random conditional injection mechanism based on physical knowledge, in which physical property descriptors are incorporated as auxiliary text prompts to refine the generative process for better consistency with specified physical properties. Third, we introduce a dual alignment strategy to enforce semantic and structural consistency further. This includes: (1) a CLIP-based contrastive encoder to align textual and structural representations during pre-training; and (2) a reward-guided sampling module that optimizes the diffusion process using a learned evaluator based on semantic and geometric alignment scores.

3.1. Diffusion-Based Text-to-Microstructure

This section introduces a diffusion-based inverse design framework that constitutes the foundation of our methodology. The framework is designed to generate 3D microstructures that align with high-level semantic descriptions while maintaining structural plausibility, thereby enabling precise and controllable mapping between structural geometry and functional performance, where each 3D microstructure is represented as a voxel grid capturing material occupancy in space. Initially, the model is based on the denoising diffusion probabilistic model (DDPM) [22], where the forward process Gaussian denoising scheme, gradually perturbs the clean samples x_0 into noisy samples x_t, defined as:

$$x_t = \sqrt{\gamma_t}\, x_0 + \sqrt{1-\gamma_t}\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and γ_t decreases steadily from 1 to 0. In the reverse process, a U-Net-like 3D network is employed to reconstruct x_0 from the noisy input x_t.

To improve reconstruction quality and stability, a self-conditional diffusion model [7] is adopted, which recursively incorporates the previous prediction of the model as an auxiliary input $\hat{x}_0'$ at each denoising step. By leveraging contextual information beyond the current noisy input, the model iteratively refines its estimations to achieve more accurate reconstruction. To mitigate excessive dependence on prior outputs and reduce cumulative errors, the auxiliary input is randomly replaced with zero during training with a fixed probability, thereby improving robustness. Specifically, the network outputs:

$$\hat{x}_0 = f(x_t, \hat{x}_0', t, z_i^d), \tag{2}$$

where $\hat{x}_0' = f(x_t, 0, t, z_i^d)$ is either the previous prediction of the model (with 50% probability) or zero. The text condition z_i^d is encoded using a pretrained text encoder, followed by sinusoidal time embeddings, and integrated into the denoising network via classifier-free guidance. The model is optimized by minimizing the reconstruction loss:

$$\mathcal{L}_{con} = E_{\epsilon,t} \left\| \hat{x}_0 - x_0 \right\|_2^2, \tag{3}$$

which encourages the alignment of text semantics with geometric structures through iterative denoising.

3.2. Property-Informed Stochastic Conditioning

To effectively incorporate physical property guidance into the inverse design process, we represent quantitative material physical property as augmented textual conditions, enabling seamless integration with semantic prompts. These

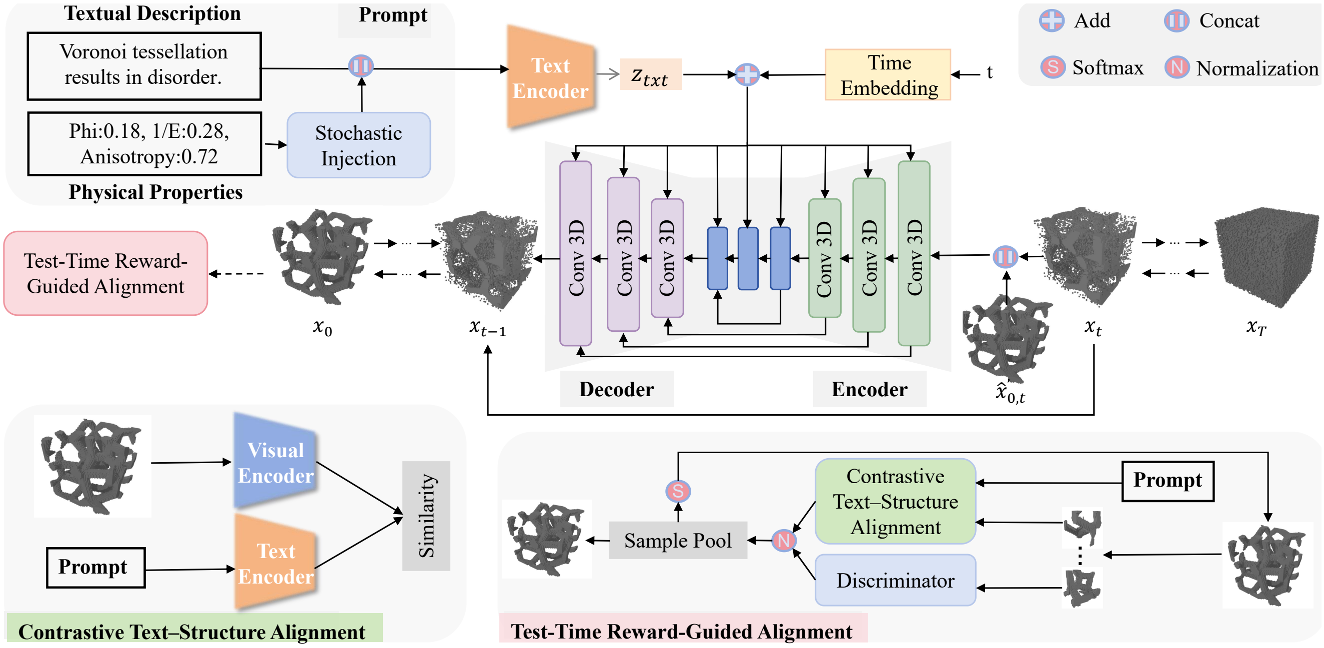


Figure 2. **Pipeline of the proposed PropDiff-TMG:** Our method is based on a self-conditional diffusion model, guided by textual descriptions and optionally injected physical properties. During training, a contrastive text–structure alignment strategy aligns text and structure representations via a contrastive loss. At inference, we further refine generations via reward-guided sampling using a CLIP and a discriminator, achieving dual alignment from both pretraining and optimization perspectives.

enriched conditions are embedded and injected into the diffusion model using a FiLM-based modulation mechanism, which adaptively modulates intermediate network features to robustly guide microstructure generation toward satisfying both semantic and physical constraints.

Physical Property Text Encoding: To enable inverse design optimized by quantitative physical properties, we augment the textual description T with explicit material properties such as Young's modulus E, isotropy index I, and volume fraction V_f, represented in descriptive textual form. The combined conditioning $\tilde{T} = \{T, E, I, V_f\}$ guides the diffusion model $\mathcal{G}$ to generate microstructures x that not only adhere to semantic constraints but are also optimized to meet the specified physical targets: $x \sim \mathcal{G}(\tilde{T})$. we apply random masking to the physical property conditioning $P = \{E, I, V_f\}$ during training. Each property $p \in \{E, I, V_f\}$ is retained independently with probability r_p, enabling the model to learn from both fully and partially specified prompts. This stochastic injection strategy allows the model to flexibly handle various conditioning scenarios, both with and without property constraints, and better align generation with physical targets. To assess the physical fidelity of generated structures, we train a regression network $\mathcal{R}$ using paired data $\{(x_i, P_i)\}_{i=1}^N$, where x_i denotes the i-th material structure and P_i is the corresponding physical properties. The trained regressor is then used to predict properties $\hat{P} = (\hat{E}, \hat{I}, \hat{V}_f)$ from a generated structure: $\hat{P} = \mathcal{R}(X)$. The prediction accuracy provides a quantitative measure of how well the generated microstructure conforms to the target physical properties embedded in the conditioning prompt.

FiLM-Based Text Embedding Integration: To further strengthen conditioning, a FiLM module [34] modulates intermediate denoising features based on the text embedding. Given noisy input x_t, let $F \in R^{C \times D \times H \times W}$ be an intermediate feature tensor and $e \in R^D$ the text embedding. FiLM applies an affine transformation:

$$F' = \gamma \cdot F + \beta, \tag{4}$$

where γ, β are generated by linear projections of e. This feature-wise modulation allows the model to adaptively optimize microstructure generation according to both semantic descriptions and quantitative physical property guidance.

3.3. Dual Alignment of 3D Structure and Text

We propose a dual alignment strategy to enhance the consistency between generated 3D microstructures and their corresponding textual descriptions and physical properties. First, two separate encoders aligns the structure and text embeddings during pre-training. Second, at test time, reward-guided sampling is used to optimize the structure based on the feedback from the learned reward model. These steps jointly enhance the fidelity of both semantics and structure.

Contrastive Text–Structure Alignment: Inspired by previous work [5, 33, 37], we employ two separate encoders [11, 44] pretrained to align textual and structural representations within a shared embedding space via contrastive learning. Let d_i denote a textual description of the i-th material and $x_i \in R^{D\times H\times W}$ be the corresponding 3D material structure. The text encoder f_θ maps d_i into a text embedding $z_i^d = f_\theta(d_i) \in R^D$, while the visual encoder g_ϕ encodes the structure x_i into a structure embedding $z_i^x = g_\phi(x_i) \in R^D$.

We adopt a symmetric contrastive alignment objective inspired by cross-modal similarity distribution matching. For each training sample, the similarity logits between modalities are computed as $S_{i,j} = \langle z_i^d, z_j^x\rangle/\tau$, where τ is a temperature parameter. To provide soft supervision targets, we further construct intra-modal similarity matrices $S_{i,j}^{dd}$, $S_{i,j}^{xx}$ for both modalities and average them to obtain a target similarity distribution. The soft targets are defined as:

$$S_{i,j}^{dd} = \langle z_i^d, z_j^d\rangle, \quad S_{i,j}^{xx} = \langle z_i^x, z_j^x\rangle, \tag{5}$$

$$T_{i,j} = \text{softmax}\left(\frac{2(S_{i,j}^{dd} + S_{i,j}^{xx})}{\tau}\right). \tag{6}$$

Based on these targets, we compute the final alignment loss in a bidirectional manner across modalities:

$$\mathcal{L}_{forward} = -\sum_{i=1}^{N}\sum_{j=1}^{N} T_{i,j}\log\text{softmax}(S_{i,j}), \tag{7}$$

$$\mathcal{L}_{backward} = -\sum_{i=1}^{N}\sum_{j=1}^{N} T_{j,i}\log\text{softmax}(S_{j,i}), \tag{8}$$

$$\mathcal{L}_{align} = \frac{1}{2N}\left(\mathcal{L}_{forward} + \mathcal{L}_{backward}\right). \tag{9}$$

This loss aligns material descriptions with 3D structures by matching cross-modal similarities to the average of intra-modal similarity patterns.

Test-Time Reward-Guided Alignment: To improve the semantic relevance and structural fidelity of generated microstructures, we propose test-time reward-guided alignment. Starting from initial diffusion-based samples, this method iteratively edits local regions using reward feedback. In each round, multiple candidates are sampled, evaluated, and the best edits are retained. Then, a soft resampling step selects the next input from the reward-weighted pool of best structures. To balance semantic consistency and structural plausibility, we design two reward models:

- **Contrastive reward:** This reward is computed as the cosine similarity between text and structure embeddings extracted by a visual and text encoders pre-training with CLIP-style text-structure alignment.
- **Discriminative reward:** This reward indicates the score reflecting the authenticity of generated microstructures. We design a discriminator network with 3D convolutional layers, batch normalization, and fully connected layers to distinguish between real and generated structures. This network is trained with binary cross-entropy loss using voxels of both real and generated 3D structures. During testing, it generates a score which represents the fine-grained structural plausibility.

To adapt to the scales of the two rewards, we design a weighted normalization to evaluate the quality of the generated structure. For the k-th candidate structure of the i-th textual prompt, the final reward score is computed as:

$$R_{i,k} = \tilde{R}_{i,k}^c + w\cdot\tilde{R}_{i,k}^d, \tag{10}$$

where w is the weight hyperparameter, $\tilde{R}_{i,k}^c$ and $\tilde{R}_{i,k}^d$ are the normalized contrastive and discriminative rewards, respectively. Normalization is carried out by subtracting the mean and dividing by the standard deviation within each batch.

4. Experiment

4.1. Datasets and Evaluation Metrics

To evaluate the text-guided diffusion model for material structure generation, we conduct comprehensive experiments using the **Geometries 2000** dataset comprising 2000 text-structure pairs of various types of materials [61].

Due to the limited scale of the Geometries 2000, we construct **GenText-Microstruct** with textual descriptions, a large text–to-microstructure dataset derived from mechanical metamaterials [53]. Descriptions are first generated by GPT based on conditions including properties, and then manually verified. The dataset contains over 14,000 samples for training and 2,000 for evaluation, covering wide ranges of modulus and Poisson's ratio. Representative structural examples are illustrated in Figure 3.

We perform a quantitative analysis using four complementary metrics: classification accuracy, Fréchet Inception Distance (FID), CLIP score, and Chamfer Distance (CD). These metrics respectively measure semantic consistency, distributional fidelity, text-to-structure alignment, and geometric accuracy of the generated microstructures.

4.2. Results of Text-to-Microstructure Generation

We conduct both qualitative and quantitative evaluations to assess the ability of model to generate microstructures under various textual conditions.

Evaluation on Geometries 2000: Table 1 shows microstructure generation results on the Geometries 2000. Our method achieves superior accuracy, suggesting that the generated structures exhibit a stronger correspondence with the intended target labels. Besides, the lower FID indicates that the generated microstructures more closely match the feature distribution of real samples.

Table 1. **Quantitative comparisons on the Geometries 2000 [61]:** The first four metrics are calculated using 200 random samples from the 2000 datasets, after setting a random seed. For comparison with Txt2Microstruct-Net, the FID score is calculated using 200 prompts that generate 10 prompts with 2000 data. R^2-square is used to assess the goodness of fit between the properties of the 2000 true structure and the predicted properties of the generated structure. Our baseline framework first aligns textual and structural semantics, and then performs direct unconstrained diffusion-based generation of 3D structures.

Method	Accuracy ↑	FID ↓	CLIP ↑	CD ↓	R^2-square ↑	Input	Method
Txt2Microstruct-Net [61]	0.8695	72.08	0.5599	0.0932	0.773 0.795 0.771	Text	VAE
Baseline	0.8959	186.54	0.5856	0.0694	0.849 0.772 0.886	Text	Diffusion
PropDiff-TMG (ours)	**0.9100**	**70.81**	**0.6936**	**0.0395**	**0.961 0.928 0.956**	Text	Diffusion

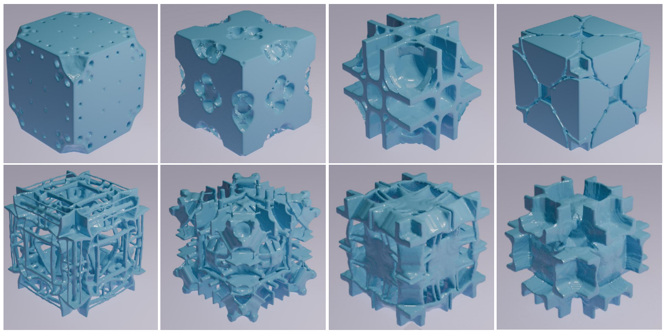


Figure 3. **Sample 3D structures from GenText-Microstruct:** The dataset comprises metamaterials spanning a broad range of physical properties, distinguishing it from Geometries 2000 [61].

Furthermore, with property-conditioned constraints, the structures exhibit more accurate predicted properties. The higher coefficient of determination and improved linear regression performance, as illustrated in Figure 5, demonstrate that our approach is more effective in inverse design that satisfies specified physical constraints. Meanwhile, property error is calculated to verify the mechanical feasibility of the generated structures, as shown in Table 2. Additionally, we incorporate the CLIP score to evaluate the quantitative alignment between input conditions and generated structures. The data shows that the generated results are semantically coherent with their descriptions.

Qualitative Analysis: Rather than relying on design templates, the model autonomously constructs diverse structures guided by natural text. As shown in Figure 4, for a single textual prompt, the model is able to generate multiple outputs with different structures, reflecting the inherent one-to-many nature of text-to-structure mapping. This generation diversity under consistent semantic constraints helps expand the material design space and promotes the discovery of novel or unconventional microstructures. Furthermore, the high consistency between textual and structural semantics allows the model to effectively capture textual semantics and improve manufacturing feasibility. In addition, visual results generated from GPT-based text prompts are provided in the supplementary material.

Table 2. **Property error results.**

Method	Young's modulus ↓	Anisotropy ↓	Volume fraction ↓
Txt2Microstruct-Net	**0.0118**	0.0163	0.0348
PropDiff-TMG (ours)	0.0175	**0.0106**	**0.0103**

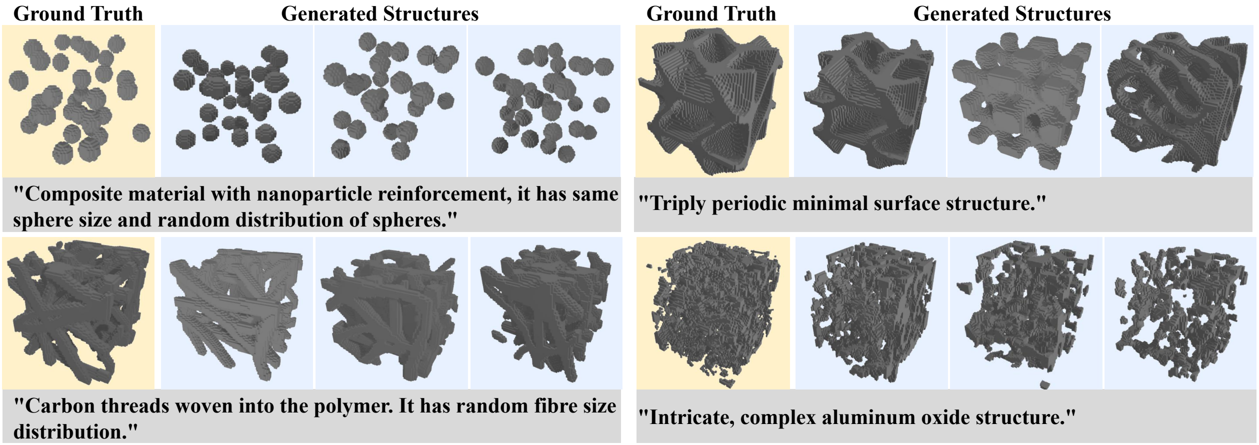
Table 3. **Quantitative comparisons on the GenText-Microstruct:** The baseline is based on properties prompt to generate structures, which follows the strategies used in previous study [53]. Reward-Guided Align denotes the module of Test-Time Reward-Guided Alignment. Since CLIP measures the semantic correspondence between the text description and the generated structure, the baseline performs poorly on this metric.

Method	FID ↓	CLIP ↑	CD ↓
Baseline [53]	84.46	0.3281	0.0666
PropDiff-TMG (ours)	**47.74**	**0.6463**	**0.0442**
w/o Property Condition	52.94	0.5210	0.0482
w/o Reward-Guided Align	49.02	0.5164	0.0468

Table 4. **Ablation studies:** Contrastive Align and Reward-Guided Align denote Contrastive Text–Structure Alignment and Test-Time Reward-Guided Alignment, respectively.

	Method	FID ↓	CLIP ↑	CD ↓
①	PropDiff-TMG (ours)	**70.81**	**0.6936**	**0.0395**
②	w/o Property Condition	105.35	0.6816	0.0651
③	w/o Contrastive Align	264.63	0.5161	0.0579
④	w/o Reward-Guided Align	81.68	0.6078	0.0412
⑤	w/o Discriminator	73.51	0.7038	0.0396
⑥	w/o Normalization	77.52	**0.7189**	**0.0394**

Evaluation on GenText-Microstruct: Table 3 showcases our performance on the GenText-Microstruct dataset. It shows that our method significantly reduces FID and CD scores relative to baseline, reflecting improvements in fidelity and distribution realism, and further demonstrating the robustness and effectiveness of the proposed framework. Furthermore, both property-informed stochastic conditioning and test-time reward-guided alignment improved the performance of each metric, demonstrating the effectiveness and robustness of each module.



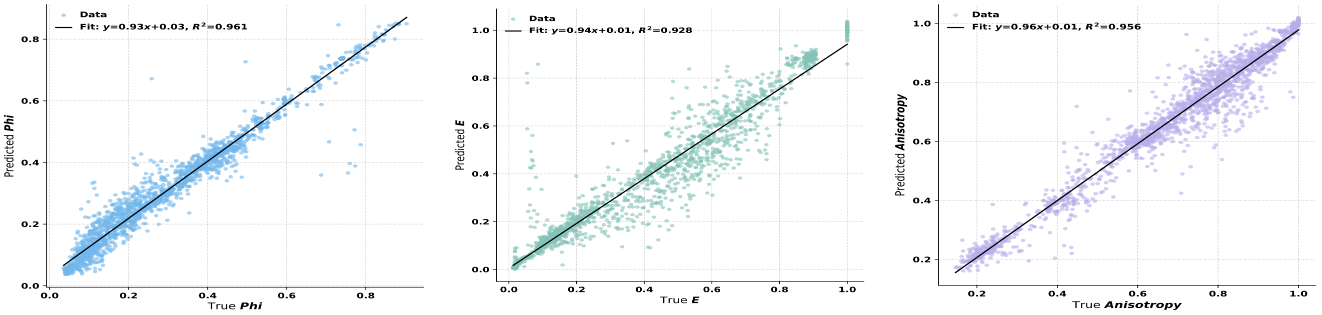
"Composite material with nanoparticle reinforcement, it has same sphere size and random distribution of spheres."

"Triply periodic minimal surface structure."

"Carbon threads woven into the polymer. It has random fibre size distribution."

"Intricate, complex aluminum oxide structure."

Figure 4. **Qualitative visualizations:** Voxel-based microstructures generated by our model using textual prompts.



(a) Volume score comparison

(b) Young's modulus comparison

(c) Anisotropy comparison

Figure 5. **Comparison of structural physical properties:** Linear regression plots between the true values and the generated physical properties for 2,000 generated microstructures with property conditions. Each subplot compares one property: (a) volume fraction, (b) Young's modulus, and (c) anisotropy score. Higher linearity and lower deviation from the fitted line indicate higher generated accuracy.

4.3. Ablation Studies

To comprehensively evaluate the contributions of different components in our method, we conduct ablation studies on different modules, as shown in Table 4.

Effectiveness of Property-Informed Stochastic Conditioning: To evaluate the effect of conditioning on physical properties, we trained the model with additional supervision using property values such as Young's modulus, volume fraction, and anisotropy. It found that even without providing any physical properties during inference, the model trained with these conditionings produced higher quality structures than the baseline model trained without these conditionings, with the CD score dropping significantly from 0.0651 to 0.0395. This suggests that incorporating physical property information during training can help guide the model to learn more physically meaningful and structurally consistent representations.

Effectiveness of Contrastive Text-Structure Alignment: To evaluate the effectiveness of contrastive text-structure alignment, we directly use the self-conditional diffusion model for training instead of pre-aligning the text and structure. It found that incorporating contrastive text-structure alignment during pre-training significantly improved performance across multiple metrics, with the FID score exhibiting a substantial reduction from 264.63 to 70.81, and the CLIP score significantly increased from 0.5161 to 0.6936. This is primarily due to the incorporation of specialized representations of the material domain during pre-training. This effectively mitigated the semantic bias of general language models in understanding material-related text and enhanced the semantic consistency between structure and description.

Effectiveness of Test-Time Reward-Guided Alignment: To evaluate the effectiveness of test-time reward-guided alignment, which refines the generated structure by iteratively updating random regions with candidate regions that have higher rewards. This strategy significantly improves various metrics, demonstrating its ability to guide the model towards the desired semantic and structural goals without additional training.

In addition, we conduct ablation experiments on each component in this module. When there is only a reward model for text-structure alignment, the CLIP metric is significantly improved to 0.7038. The normalized combined

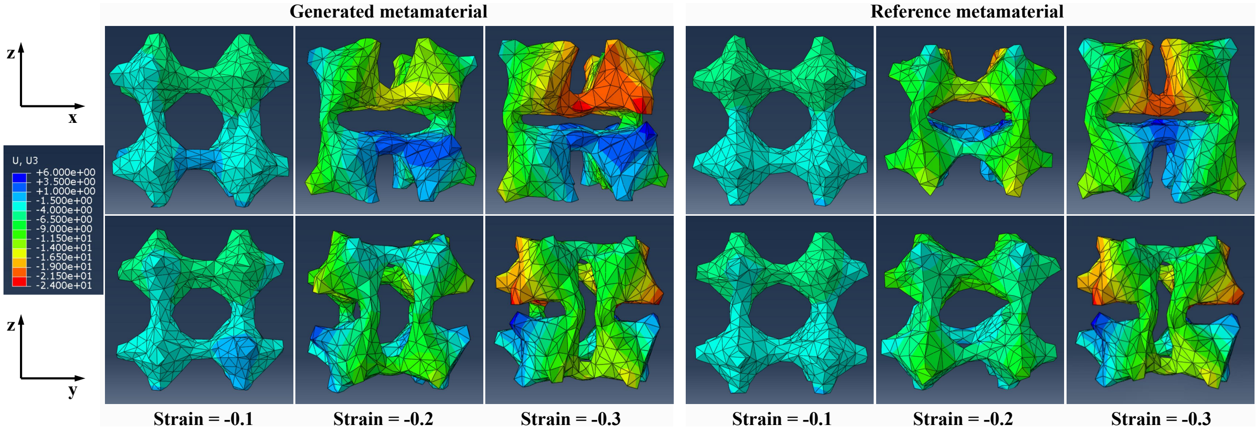
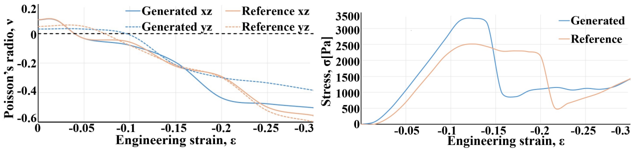


Figure 6. **Qualitative results of simulations:** Qualitative comparison of mechanical properties of generated metamaterial with those of the reference. Gradual deformed shapes at different compressive strains obtained from finite element simulations.



(a) Poisson's-strain curve

(b) Stress-strain curve

Figure 7. **Mechanical properties by simulation:** Quantitative comparison of (a) Poisson's ratio strain curve and (b) stress-strain curve between the generated material and the reference material obtained from finite element simulation.

reward model, after adding the discriminator reward model, not only improves the CLIP metric to 0.6936, but also significantly reduces the FID score to 70.81. This shows that the reward model for text-structure alignment acts on semantic consistency, while the discriminator reward model acts on structural rationality.

4.4. Finite Element Method Simulation

Finite element method simulations are performed using ABAQUS to investigate the large-deformation behavior of the generated metamaterial structure and a reference metamaterial. The large-deformation response of the auxetic metamaterial is investigated using a geometrically nonlinear static analysis. The material is modeled using a nearly incompressible neo-Hookean hyperelastic model with a Young's modulus set to 0.6615 MPa. To capture buckling-related behavior, a linear eigenvalue buckling analysis is first performed. subsequently, a postbuckling simulation is performed, gradually compressing the structure along the axial direction (z) by a prescribed displacement, terminating the analysis when contact is detected between opposing boundaries. The postbuckling stress-strain response and the evolution of the Poisson's ratio are extracted by tracking the reaction forces and displacements on the loaded surface and the lateral surfaces, respectively.

Figure 6 shows a sequence of the progressively deformed metamaterial generated under four different levels of compressive engineering stress, obtained from the finite element results. Compared with the reference material, the generated metamaterial exhibits favorable negative Poisson ratio properties in both the x- and y-directions, although the deformation process is relatively distorted.

Figure 7 provides a detailed comparison of the Poisson's ratio-strain and stress-strain curves obtained from simulations. The results show that the Poisson's ratio of the generated structure decreases with increasing compressive strain. Some fluctuations in the curve may be due to the simulation setup, which also affects the stress-strain curve. Although the generated structure may have defects that cause stress to drop, it still exhibits a mechanical trend similar to that of the reference structure during deformation. Experimental results demonstrate that the generative model based on the diffusion model can produce structures that are consistent with the semantic and physical properties targets.

5. Conclusion

In this work, we propose PropDiff-TMG, a unified diffusion framework for inverse design of 3D metamaterial microstructures from textual descriptions and physical properties. To optimize the microstructure, we introduce quantitative properties in addition to feature description. Furthermore, a dual alignment strategy is proposed to eliminate cross-domain knowledge differences through text–structure alignment and to strengthen the alignment of semantic and physical structures through a test-time reward-guided alignment. Extensive experiments show that PropDiff-TMG is able to generate diverse, physically plausible, and semantically consistent microstructures, achieving moderate improvement over existing microstructure methods.

Acknowledgments

This work was supported by National Science Foundation of China (52441503, 62302093), Jiangsu Province Natural Science Fund (BK20230833), the CIPS-SMP-Zhipu Large Model Fund, the Fundamental Research Funds for the Central Universities (2242025K30024), the Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (E5SP060116) and the Big Data Computing Center of Southeast University.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12608–12618, 2023. 3

[2] Jan-Hendrik Bastek and Dennis M Kochmann. Inverse design of nonlinear mechanical metamaterials via video denoising diffusion models. *Nature Machine Intelligence*, 5 (12):1466–1475, 2023. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[4] Markus J Buehler. Generating 3d architectured nature-inspired materials and granular media using diffusion models based on language cues. *Oxford Open Materials Science*, 2 (1):itac010, 2022. 2

[5] Lianggangxu Chen, Xuejiao Wang, Jiale Lu, Shaohui Lin, Changbo Wang, and Gaoqi He. Clip-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27863–27873, 2024. 5

[6] Shuai Chen, Zachary H Aitken, Subrahmanyam Pattamatta, Zhaoxuan Wu, Zhi Gen Yu, David J Srolovitz, Peter K Liaw, and Yong-Wei Zhang. Simultaneously enhancing the ultimate strength and ductility of high-entropy alloys via short-range ordering. *Nature Communications*, 12(1):4953, 2021. 1

[7] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 3

[8] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26576–26586, 2025. 3

[9] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022. 2

[10] Kishalay Das, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. *arXiv preprint arXiv:2503.00522*, 2025. 3

[11] 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. 5

[12] Simon Divilov, Hagen Eckert, David Hicks, Corey Oses, Cormac Toher, Rico Friedrich, Marco Esters, Michael J Mehl, Adam C Zettel, Yoav Lederer, et al. Disordered enthalpy–entropy descriptor for high-entropy ceramics discovery. *Nature*, 625(7993):66–73, 2024. 1

[13] Zongliang Du, Xiao-Yi Zhou, Renato Picelli, and H Alicia Kim. Connecting microstructures for multiscale topology optimization with connectivity index constraints. *Journal of Mechanical Design*, 140(11):111417, 2018. 2

[14] Andreas Engelhardt, Mark Boss, Vikram Voleti, Chun-Han Yao, Hendrik Lensch, and Varun Jampani. Svim3d: Stable video material diffusion for single image 3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28428–28439, 2025. 3

[15] Jie Gao, Mi Xiao, Yan Zhang, and Liang Gao. A comprehensive review of isogeometric topology optimization: methods, applications and prospects. *Chinese Journal of Mechanical Engineering*, 33:1–14, 2020. 2

[16] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3

[17] Zhenyang Gao, Pengyuan Ren, Yifeng Dong, Gengchen Zheng, Min-Son Pham, Xiao Shang, Shaojia Wang, Shuo Yang, Zijue Tang, Yongbing Li, et al. Polymer-inspired mechanical metamaterials. *arXiv preprint arXiv:2512.16732*, 2025. 1

[18] Zhenyang Gao, Gengchen Zheng, Pengyuan Ren, Hongsong Wang, Kun Zhou, Minh-Son Pham, Yi Wu, Yu Zou, Chu Lun Alex Leung, Yuanyuan Tian, et al. Generative metamaterials based on large language models. *arXiv preprint arXiv:2601.17997*, 2026. 1, 2

[19] Eric Garner, Helena MA Kolken, Charlie CL Wang, Amir A Zadpoor, and Jun Wu. Compatibility in microstructural optimization for additive manufacturing. *Additive Manufacturing*, 26:65–75, 2019. 2

[20] Jiacheng Han, Xiaoya Zhai, Lili Wang, Di Zhang, Junhao Ding, Winston Wai Shing Ma, Xu Song, Wei-Hsin Liao, Ligang Liu, Jun Wu, et al. Inverse-designed 3d sequential metamaterials achieving extreme stiffness. *Materials & Design*, 247:113350, 2024. 2

[21] Zebin He, Mingxin Yang, Shuhui Yang, Yixuan Tang, Tao Wang, Kaihao Zhang, Guanying Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Materialmvp: Illumination-

invariant material generation via multi-view pbr diffusion. *arXiv preprint arXiv:2503.10289*, 2025. 3

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[23] Yu-Chuan Hsu, Zhenze Yang, and Markus J Buehler. Generative design, manufacturing, and molecular modeling of 3d architected materials based on natural language input. *APL Materials*, 10(4), 2022. 2

[24] Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26556–26565, 2025. 3

[25] Pengcheng Jiao and Amir H Alavi. Artificial intelligence-enabled smart mechanical metamaterials: advent and future trends. *International Materials Reviews*, 66(6):365–393, 2021. 2

[26] Hunter T Kollmann, Diab W Abueidda, Seid Koric, Erman Guleryuz, and Nahil A Sobh. Deep learning for topology optimization of 2d metamaterials. *Materials & Design*, 196: 109098, 2020. 2

[27] Siddhant Kumar, Stephanie Tan, Li Zheng, and Dennis M Kochmann. Inverse-designed spinodoid metamaterials. *npj Computational Materials*, 6(1):73, 2020. 2

[28] Doksoo Lee, Wei Chen, Liwei Wang, Yu-Chin Chan, and Wei Chen. Data-driven design for metamaterials and multiscale systems: a review. *Advanced Materials*, 36(8): 2305254, 2024. 2

[29] Junhyeong Lee, Donggeun Park, Mingyu Lee, Hugon Lee, Kundo Park, Ikjin Lee, and Seunghwa Ryu. Machine learning-based inverse design methods considering data characteristics and design space size in materials design and manufacturing: a review. *Materials horizons*, 10(12):5436–5456, 2023. 2

[30] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26069–26080, 2025. 3

[31] Kiara L McMillan, Doğacan S Öztürk, and Pinar Acar. Inverse design of 2d-mechanical metamaterials with spinodal topologies under uncertainty. In *AIAA SCITECH 2022 Forum*, page 0811, 2022. 1

[32] Mehrdad Mohsenizadeh, Federico Gasbarri, Michael Munther, Ali Beheshti, and Keivan Davami. Additively-manufactured lightweight metamaterials for energy absorption. *Materials & Design*, 139:521–530, 2018. 1

[33] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5607–5612, 2023. 5

[34] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2, 4

[35] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9914–9925, 2024. 3

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 5

[38] Minsik Seo and Seungjae Min. Dl-msto+: A deep learning-based multi-scale topology optimization framework via positive definiteness ensured material representation network. *Computer Methods in Applied Mechanics and Engineering*, 415:116276, 2023. 2

[39] Wei Sha, Mi Xiao, Yihui Wang, Mingzhe Huang, Qishi Li, and Liang Gao. Topology optimization methods for thermal metamaterials: A review. *International Journal of Heat and Mass Transfer*, 227:125588, 2024. 1

[40] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24130–24141, 2024. 3

[41] Ole Sigmund. Systematic design of metamaterials by topology optimization. In *IUTAM Symposium on Modelling Nanomaterials and Nanosystems: Proceedings of the IUTAM Symposium held in Aalborg, Denmark, 19–22 May 2008*, pages 151–159. Springer, 2009. 2

[42] JunHo Song, JaeHoon Lee, Namjung Kim, and Kyoungmin Min. Artificial intelligence in the design of innovative metamaterials: A comprehensive review. *International Journal of Precision Engineering and Manufacturing*, 25(1):225–244, 2024. 2

[43] P Suhas, Jaimon Dennis Quadros, Yakub Iqbal Mogul, Ma Mohin, Abdul Aabid, Muneer Baig, and Omar Shabbir Ahmed. A review of mechanical metamaterials and additively manufacturing techniques for biomedical applications. *Materials Advances*, 2025. 1

[44] 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. 5

[45] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 2

[46] Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. Matfuse: controllable material generation with diffusion models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4429–4438, 2024. 3

[47] Chen Wang, Hao-Yang Peng, Ying-Tian Liu, Jiatao Gu, and Shi-Min Hu. Diffusion models for 3d generation: A survey. *Computational Visual Media*, 11(1):1–28, 2025. 3

[48] Haotian Wang, Chen Pan, Haiyuan Zhao, Tingyu Wang, and Yafeng Han. Design of a metamaterial film with excellent conformability and adhesion for bandage substrates. *Journal of the mechanical behavior of biomedical materials*, 124: 104799, 2021. 2

[49] Jia Wang, Yingxue Wang, and Yanan Chen. Inverse design of materials by machine learning. *Materials*, 15(5):1811, 2022. 1

[50] Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, and Kim-Hui Yap. Taps3d: Text-guided 3d textured shape generation from pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16805–16815, 2023. 2

[51] Jun Wu, Ole Sigmund, and Jeroen P Groen. Topology optimization of multi-scale structures: a review. *Structural and Multidisciplinary Optimization*, 63(3):1455–1480, 2021. 1

[52] Yuliang Xia, Yang He, Fenghua Zhang, Yanju Liu, and Jinsong Leng. A review of shape memory polymers and composites: mechanisms, materials, and applications. *Advanced Materials*, 33(6):2000713, 2021. 1

[53] 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. 2, 5, 6

[54] Zhenze Yang and Markus J Buehler. Words to matter: De novo architected materials design using transformer neural networks. *Frontiers in Materials*, 8:740754, 2021. 2

[55] Ee Teng Zhang, Hu Liu, and Bing Feng Ng. Mechanics of re-entrant anti-trichiral honeycombs with nature-inspired gradient distributions. *International Journal of Mechanical Sciences*, 259:108597, 2023. 2

[56] Yuhong Zhao, Hui Xing, Lijun Zhang, Houbing Huang, Dongke Sun, Xianglei Dong, Yongxing Shen, and Jincheng Wang. Development of phase-field modeling in materials science in china: a review. *Acta Metallurgica Sinica (English Letters)*, 36(11):1749–1775, 2023. 1

[57] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems*, 36:73969–73982, 2023. 2

[58] Li Zheng, Siddhant Kumar, and Dennis M Kochmann. Diffumeta: Algebraic language models for inverse design of metamaterials via diffusion transformers. *arXiv preprint arXiv:2507.15753*, 2025. 3

[59] Xiaoyang Zheng, Xiaofeng Guo, and Ikumu Watanabe. A mathematically defined 3d auxetic metamaterial with tunable mechanical and conduction properties. *Materials & Design*, 198:109313, 2021. 2

[60] Xiaoyang Zheng, Ta-Te Chen, Xiaoyu Jiang, Masanobu Naito, and Ikumu Watanabe. Deep-learning-based inverse design of three-dimensional architected cellular materials with the target porosity and stiffness using voxelized voronoi lattices. *Advanced Materials*, 24(1):2157682, 2023. 1, 2

[61] 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. 2, 5, 6

[62] Jiaxin Zhou, Ikumu Watanabe, and Takayuki Yamada. Computational morphology design of duplex structure considering interface debonding. *Composite Structures*, 302:116200, 2022. 2