TUNING MULTIMODAL LARGE LANGUAGE MODEL FOR MICROSTRUCTURE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although 3D material microstructure generation has achieved notable progress, existing works typically rely on domain-specific constraints, leaving research oriented to natural languages largely unexplored. Inspired by the recent success of Large Language Models (LLMs) across diverse domains, this work introduces a multi-conditional generation framework that supports text-, sketch-, and property-driven microstructure design. The proposed method employs a Vector Quantized Variational AutoEncoder (VQ-VAE) to encode microstructural representations and a visual encoder to process sketches, allowing multi-modal alignment between text, sketch and structure within an LLM-based architecture. Moreover, to ensure that the generated microstructures satisfy desired physical properties, property constraints are incorporated into the optimization process. Experimental results demonstrate that the proposed framework can produce high-quality and semantically consistent microstructures, suggesting its potential to support more intuitive and accessible microstructure design workflows.

1 INTRODUCTION

Metamaterials (Wu et al., 2026; Chen et al., 2021; Xia et al., 2021; Divilov et al., 2024) are a class of artificially engineered material systems composed of periodic or quasi-periodic microstructural units, whose macroscopic physical properties are predominantly governed by geometric configurations rather than chemical composition. By tailoring the topology and length scales of the microstructures, metamaterials can exhibit unconventional mechanical (Berger et al., 2017), acoustic (Noguchi et al., 2022), electromagnetic (Ma et al., 2022) and thermal behaviors (Xiao et al., 2023) that are difficult to achieve with natural materials. Consequently, microstructural inverse design for target properties has emerged as a key research direction in advanced materials development.

With the rapid advancement of deep learning, material inverse design has progressively shifted from conventional computational approaches, such as phase-field simulations (Zhao et al., 2023a) and topology optimization (Sha et al., 2024), toward data-driven generative models (Choudhary et al., 2022; Seo & Min, 2023). However, these approaches typically depend on highly specialized physical parameters or engineering priors as conditioning inputs, thereby limiting their accessibility for non-expert users and applicability in general design scenarios. Existing multimodal generation methods (Lee & Ahmed, 2026; Zheng et al., 2024) based on diffusion and variational autoencoder frameworks, despite enabling synthesis of 3D structures from natural language and 2D images, suffer from limited semantic comprehension and limited generalization ability.

By contrast, Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) exhibit strong capabilities in complex semantic modeling and cross-task generalization, offering a promising paradigm for multimodal condition-driven generation. While LLMs have achieved substantial success in text-based and multi-task learning (Koh et al., 2023; Hong et al., 2023), their potential in material microstructure generation, remains largely unexplored. Owing to their strengths in multimodal semantic alignment and conditional reasoning, integrating LLMs into the joint modeling of text, images, and 3D material structures may enable a more general, flexible, and user-friendly framework for material inverse design.

To achieve inverse design of material microstructures under multimodal conditions, we propose T2M-LLM, shown in Figure 1, a generative framework that combines text descriptions, hand-drawn sketches and discrete 3D microstructure representations. The core component employs pretrained

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

LLMs as a conditional generator with lightweight adapter tuning, obviating the need for large-scale training from scratch.

In this framework, a pair of microstructure tokenizers and detokenizers is first trained to establish a bidirectional mapping between continuous microstructure representations and discrete structural tokens. By augmenting the vocabulary of a pretrained LLM with microstructure tokens and applying lightweight fine-tuning without altering the original text tokens, the LLM is enabled to synthesize microstructure token sequences conditioned on textual descriptions and hand-drawn sketches, which are subsequently decoded into final microstructures for multimodal inverse design. Although text and sketch conditions impose effective constraints on global morphology, such high-level cues are generally insufficient to ensure accurate compliance with target physical properties. Accordingly, a property-driven test-time adaptation strategy is employed during inference, wherein gradient-guided refinement of latent representations is performed to achieve closer adherence to the specified properties without compromising structural fidelity.

Collectively, these components constitute a multimodal condition-driven microstructure generation framework that enables the synthesis of high-quality, diverse, and physically plausible 3D microstructures conditioned on descriptions of natural language and hand-drawn sketches. To facilitate comprehensive training and evaluation, we construct two datasets comprising paired text descriptions, 2D sketches, and voxelize microstructures spanning multiple material categories and conduct systematic experimental studies. Extensive results demonstrate that the proposed approach achieves strong performance in both semantic alignment and consistency with target property specifications.

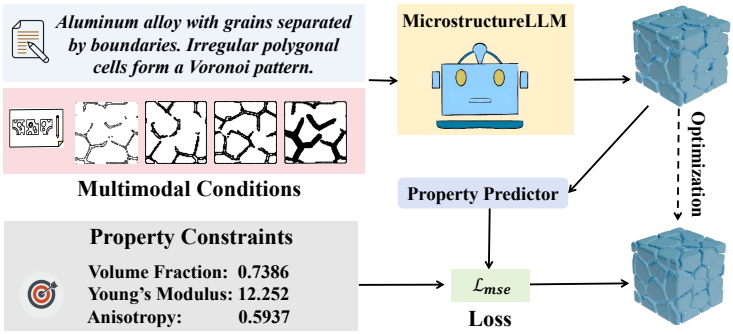


Figure 1: **Overview of our approach:** An LLM-based framework can generate metamaterial microstructures based on textual descriptions, hand-drawn sketches, and physical properties.

Contributions: 1) Multimodal material inverse design paradigm: We propose a multimodal microstructure inverse design framework to synthesize 3D microstructures from textual descriptions and hand-drawn sketches. **2) Property-guided adaptive optimization:** We introduce a property-guided test-time adaptation strategy to refine generated structures toward improved consistency with target physical properties. **3) Benchmark for text-to-microstructure:** We extend existing datasets and establish a large-scale benchmark for text-to-microstructure generation.

2 Related Work

Inverse Design of Microstructure: With the successful development of data-driven microstructure generation techniques based on generative artificial intelligence (Jiao & Alavi, 2021; Song et al., 2024; Lee et al., 2024; Gao et al., 2026), increasing attention has been devoted to the inverse design of microstructures. Recent studies have investigated material microstructure design under various conditional inputs. The majority of existing works focus on property-driven inverse design, such as diffusion-based approaches (Yang et al., 2024; Bastek & Kochmann, 2023; Zheng et al., 2023; Zhan et al., 2025) that enable end-to-end microstructure synthesis conditioned on target mechanical properties, and VAE-based approaches (Tran et al., 2025; Zhao et al., 2025) generate lattice structures or geometric parameters with desired mechanical properties using various VAE variants. In addition, MetamatBench (Chen et al., 2025a) provides unified structural representations to facilitate property-oriented material learning.

In contrast, material design conditioned on visual or semantic information has received comparatively limited attention. Some studies (Phan et al., 2024; Lee & Ahmed, 2026) explore the reconstruction of 3D structures from 2D microstructure images. Other works (Hsu et al., 2022; Buehler, 2022) generate 2D representations based on text descriptions and then convert them into 3D structural representations.

The end-to-end framework Text2Microstruct-Net (Zheng et al., 2024) uses Variational AutoEncoders (VAEs) to generate 3D microstructures. Given its high interpretability and expressive flexibility, text-to-3D microstructure design remains a promising yet underexplored research direction. Accordingly, we integrate textual descriptions, structural sketches, and property conditions to perform end-to-end material inverse design.

Multimodal Conditional Generation: Multimodal conditional image generation (Chen et al., 2025b; Hu et al., 2024) has made significant progress recently. These approaches typically employ text, images, or their combinations as conditioning signals, and constrain the output distribution of generative models through cross-modal representation learning. For complex conditioning scenarios (Zhang et al., 2023; Kim et al., 2023), prior work has explored the collaborative design of multimodal encoders and diffusion models, enabling unified modeling of semantic constraints from textual descriptions and visual entities, for example by encoding interleaved text and image conditions into a semantically aligned conditional representation (Li et al., 2024). Beyond image synthesis, this paradigm has been successfully extended to 3D object generation (Zhao et al., 2023b; Cao et al., 2026). SDFusion (Cheng et al., 2023) integrates text, images, and some shape conditions into a latent space using a multimodal encoder and cross-attention, and generates 3D shapes using a diffusion model, enabling interactive generation guided by multiple conditions. In recent materials research, CrossMatAgent (Tian et al., 2025) explored a multimodal metamaterial generation framework that combines large-scale language models with textual and image guidance to enable interactive design generation through multi-agent collaboration. In contrast, our work unifies text, hand-drawn sketches, and material structures within an LLM, enabling more controllable microstructure generation.

3 METHOD

The proposed Text-to-Microstructure Large Language Model (T2M-LLM) is a unified LLM-based framework for inverse design of 3D metamaterial microstructures conditioned on natural language descriptions, hand-drawn sketches, and target property constraints, as illustrated in Figure 2. The method first discretizes the structure using VQ-VAE and fine-tunes the LLM to achieve text-to-structure generation; then, sketch conditions are introduced to improve visual consistency; finally, target property constraints are combined to ensure that the generated structure meets the expected physical properties.

3.1 TEXT-TO-MICROSTRUCTURE VIA MLLM

To better align with the next-token prediction mechanism of LLMs, we discretize structures into token representations using VQ-VAE. In structural tokenization process, the structure is represented as $\mathbf{s} \in \mathbb{R}^{D\times W\times H}$, where D, W, H are the resolutions of the input structure. The structure is first encoded by an encoder $\mathcal{E}_s$ into latent embeddings $\mathbf{z} \in \mathbb{R}^{D/N\times W/N\times H/N\times d}$, where N is the downsampling rate and d is the latent dimensionality. Then, the flattened latent embeddings are quantized using the codebook $C = \{\mathbf{c}_k\}_{k=1}^{K}$, where K is the codebook size and each $\mathbf{c}_k \in \mathbb{R}^d$. The quantized representations $\hat{\mathbf{z}}$ are obtained as:

$$\hat{\mathbf{z}}_t = \arg\min_{\mathbf{c}_k\in C} \|\mathbf{z}_t - \mathbf{c}_k\|_2, \tag{1}$$

where t ranges from 1 to $D/N \times W/N \times H/N$. The original structure is reconstructed by a decoder $\mathcal{D}_s$ as $\hat{\mathbf{s}} = \mathcal{D}_s(\hat{\mathbf{z}})$. Following Yan et al. (2021), we optimize the VQ-VAE using a reconstruction loss and a commitment loss:

$$\mathcal{L}_{\text{vq}} = \beta \underbrace{\|\mathbf{s} - \hat{\mathbf{s}}\|_1}_{\mathcal{L}_{\text{re}}} + \underbrace{\|\mathbf{z} - \text{sg}[\hat{\mathbf{z}}]\|_2}_{\mathcal{L}_{\text{commit}}}, \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, and β are weighting factors. Besides, the codebooks are trained using an Exponential Moving Average (EMA) strategy.

To enable the LLM to understand discrete structural token sequences, we construct a structural token vocabulary based on the indices of the codebook entries and integrate it into the original text vocabulary as newly added tokens (Jiang et al., 2023). Special tokens $< Structure >$ and $< /Structure >$ are further introduced to indicate the start and end of structure generation. After extending the vocabulary, the LLM is able to jointly model the semantic information of both text and structure, and predicts the next token in an autoregressive manner by maximizing the conditional

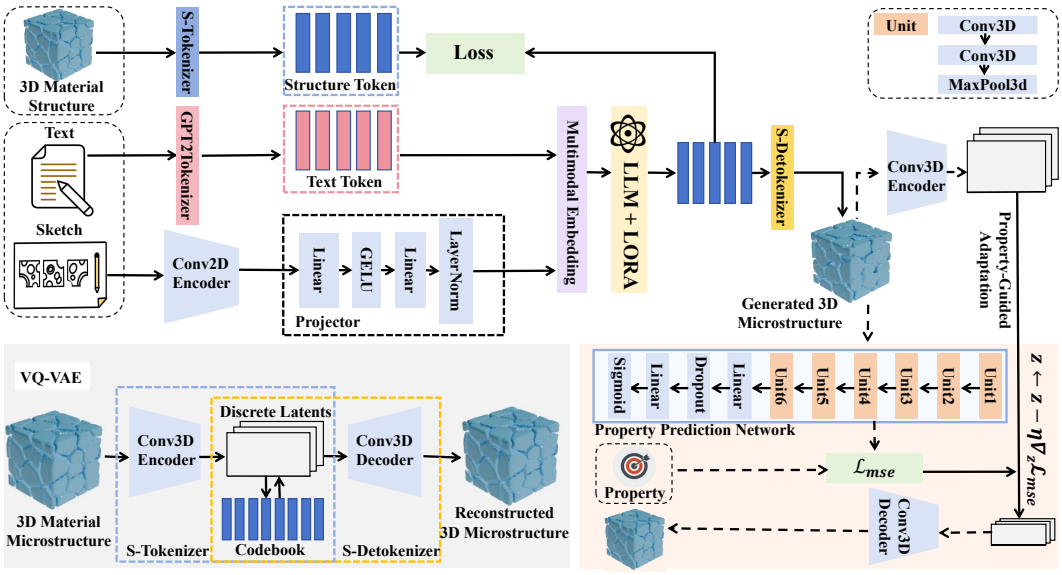


Figure 2: **Pipeline of the proposed T2M-LLM:** This approach fine-tunes a pretrained LLM for metamaterial inverse design, enabling the generation of corresponding microstructures from natural language descriptions. Hand-drawn sketches are incorporated as auxiliary visual conditions to enhance the visual consistency of the generated structures. During the inference stage, physical property constraints are further imposed to optimize the generated designs, thereby improving controllability and reliability.

probability $p_\theta(x_t \mid x_{<t}, Text)$, where $x_{1:t}$ denotes the target token sequence and $Text$ represents the caption. Accordingly, the training objective is to maximize the log-likelihood $\mathcal{L}_{\text{llm}} = -\sum \log p_\theta(x_t \mid x_{<t}, Text)$. For efficient fine-tuning, we adopt LoRA (Hu et al., 2022) and train only the additional adapters associated with the newly introduced structural tokens, significantly reducing the number of trainable parameters.

3.2 Multimodal Features in MLLM

Natural language descriptions are often insufficient for precisely conveying high-level semantics and fine-grained spatial constraints, especially for complex geometric structures. In contrast, sketches provide intuitive visual priors that better guide and constrain the generation process. To address the lack of sketch annotations in existing datasets, we propose a Structure Sketch Generation Algorithm (SSGA) based on the three-view representations of material microstructures, enabling the automatic generation of hand-drawn-style sketches. Considering the spatial nature of 3D structures, we adopt three-view projections as the input for sketch generation to provide multi-view visual priors and enhance spatial consistency. During the sketch synthesis process, voxel contours closer to the surface are rendered with thicker and darker strokes to preserve the depth-aware structural hierarchy of the 3D geometry. To emulate human hand-drawing characteristics, Gaussian blurring is applied to simulate natural stroke transitions, while random noise injection and jittered stroke rendering are introduced to model the inherent irregularity and hand tremor in manual sketches. Figure 3 illustrates the three-view projections of different material microstructures and their corresponding generated hand-drawn sketches.

To achieve joint modeling of textual descriptions, 2D sketches, and 3D microstructures, we propose a method that integrates multimodal embeddings into a sequential Transformer framework. To more effectively guide the generative process, the model is provided with instructive prompts that encourage the generation of microstructures consistent with the input descriptions and views. Specifically, we first construct an input sequence, replacing image portions with placeholders, and then automatically convert the token sequence into multimodal embedding sequence m using functions within the LLM. Subsequently, the sketches V are mapped to continuous feature representations via a visual encoder

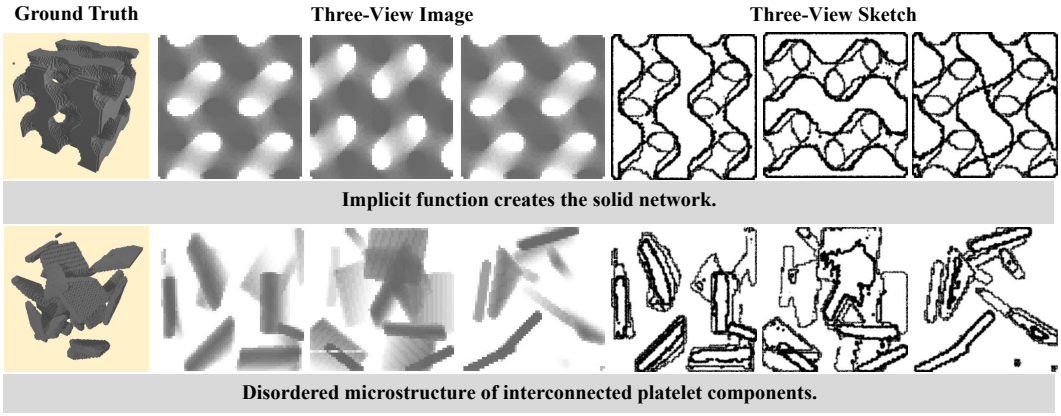


Figure 3: **Sketches generated by Structure Sketch Generation Algorithm:** It shows the 3D microstructure, three-view diagrams, and generated sketches corresponding to different texts.

$\mathcal{E}_v$, and a projector maps $\mathcal{P}_v$ these features to the dimensionality of the LLM, defined as:

$$\mathbf{v}_k = \mathcal{E}_v(V), \quad k = 1, \ldots, K, \tag{3}$$

$$\hat{\mathbf{v}}_k = \mathcal{P}_v(\mathbf{v}_k), \quad k = 1, \ldots, K. \tag{4}$$

The visual embeddings are then sequentially fused with the text embeddings at a predefined insertion position p to form the joint sequence:

$$\mathbf{x} = [\mathbf{m}_1, \ldots, \mathbf{m}_{p-1}, \mathbf{v}_1, \ldots, \mathbf{v}_K, \mathbf{m}_{p+1}, \ldots, \mathbf{m}_L]. \tag{5}$$

The model is trained end-to-end by maximizing the conditional probability of the target structural sequence given the textual and multi-view conditions:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{L_S} \log P_\theta\Big(s_i \,\Big|\, \mathbf{x}_{<i}, Text, V\Big), \tag{6}$$

where $S = \{s_1, \ldots, s_{L_S}\}$ denotes the target structural sequence, $\mathbf{x}_{<i}$ represents all previously generated embeddings in the joint sequence, $Text$ is the textual description, and V denotes the embedding(s) of either a single sketch or multiple sketches from different viewpoints. This conditional modeling design enables the model to tightly align textual and visual information while sequentially generating the microstructure, enhancing its ability to capture spatial-semantic correlations and dynamic structural patterns, thereby achieving superior performance in cross-modal generation and understanding tasks.

3.3 Property-Guided Adaptation

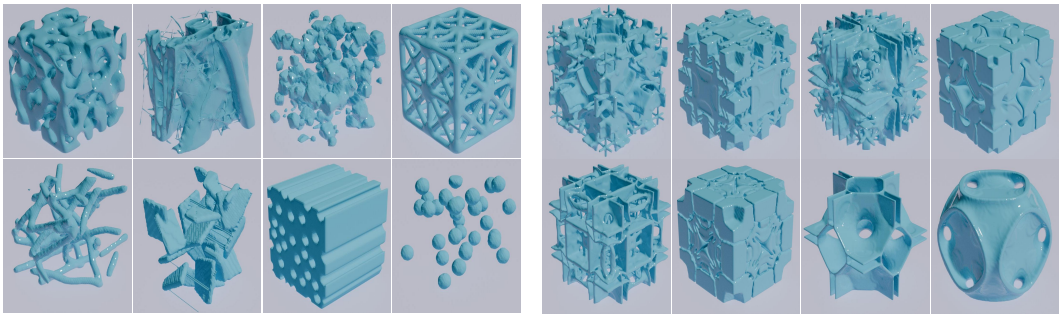
In the design of metamaterial microstructures, unique physical properties directly determine the functional performance of material. Accordingly, a property-guided adaptation auxiliary optimization strategy is introduced based on the latent variables of preliminary microstructures generated from language descriptions and visual priors. Specifically, the token sequences produced by MLLM with multimodal feature are mapped back to the VQ-VAE latent space to obtain a latent representation z_0 of the microstructure. The latent variable is then decoded through the VQ-VAE decoder $\mathcal{D}_s$ to produce a voxelized microstructure, and a pretrained physical property predictor is used to estimate key performance metrics, denoted as:

$$z = \mathcal{D}_s(z_0), \text{pred} = \mathcal{P}_{prop}(z). \tag{7}$$

To gently steer the microstructure towards user-specified target properties target without significantly altering the original semantic content, we define an auxiliary loss:

$$\mathcal{L}(z) = \|\text{pred} - \text{target}\|^2 + \lambda \|z - z_0\|^2, \tag{8}$$

where the first term measures the deviation of the predicted properties from the target values, and the second term serves as a regularization to constrain the latent variable from drifting too far from



(a) **Geometries 2000 dataset** (b) **GenText-Microstruct dataset**

Figure 4: **Sampled 3D structures from two different datasets:** (a) The Geometries 2000 includes samples of metals and alloys, ceramics, polymers and composites, and metamaterials. (b) The GenText-Microstruct includes samples of metamaterial structures covering a wide range of physics.

the initial generated representation, with λ controlling the relative weight. The latent variable z is iteratively updated using a gradient-based optimization algorithm, allowing the microstructure to gradually approach the desired property values while largely preserving the original semantic features. Since this optimization serves only as an auxiliary adjustment rather than the primary driving force for generation, the overall quality of the final microstructure may decrease slightly. Nonetheless, this strategy can still provide a feasible auxiliary optimization for more intuitive and accessible metamaterial design workflows.

4 Experiment

4.1 Experimental Setup

Datasets: To evaluate the performance of our method in material structure generation, we perform extensive experiments using an automated pipeline to construct sketch-based conditions from the **Geometries 2000** dataset (Zheng et al., 2024), which contains 2,000 text–structure pairs representing various material categories. The dataset is split into an 8:2 ratio based on a random seed of 42.

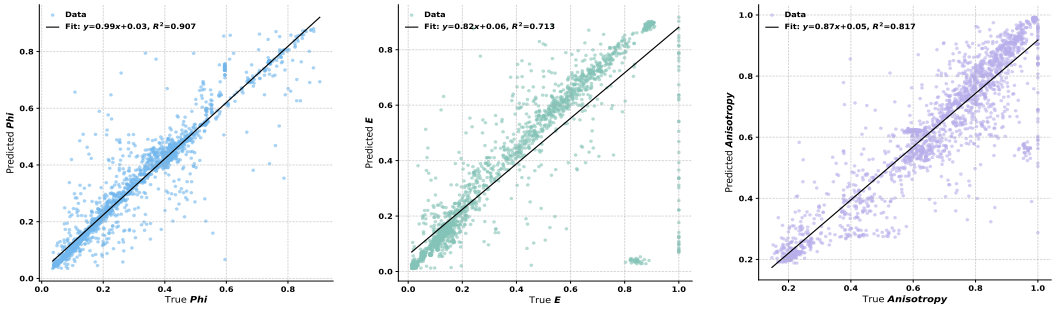
We further construct a large-scale dataset, **GenText-Microstruct**, tailored for text-driven microstructure generation based on mechanical metamaterials (Yang et al., 2024). This dataset contains over 14,000 training samples and 2,000 evaluation samples, encompassing a broad range of elastic moduli and Poisson's ratios. The textual descriptions are initially generated using a GPT-based model conditioned on material properties and subsequently refined through manual verification. Similarly, the sketch conditions are generated by applying the same automated procedure to the three orthographic views of each structure. Representative structural examples are illustrated in Figure 4.

Evaluation Metrics: For quantitative evaluation, we adopt four complementary metrics: classification accuracy, Fréchet Inception Distance (FID), CLIP score, and Chamfer Distance (CD). These metrics assess semantic consistency, distribution fidelity, text–structure alignment, and geometric accuracy of the generated microstructures, respectively. Classification accuracy is calculated using a convolutional-based classifier trained on Geometries 2000. The CLIP score is calculated as cosine similarity using a CLIP-style dual encoder trained on Geometries 2000. The FID is calculated on the basis of features extracted from the visual encoder in the dual encoder.

Implementation Details: The VQ-VAE architecture is configured with a downsampling factor $N = 8$, a hidden dimension $d = 768$, and a codebook size $K = 256$. The input microstructures have a resolution of $64 \times 64 \times 64$ voxels. GPT-2 (Radford et al., 2019) is adopted as the generative LLM to enhance the understanding of both textual and sketch-based conditions, thereby improving the model's generalization capability. The LoRA adaptation is configured with a rank (r) and scaling factor (α) of 256 for generation. Besides, we use the L-BFGS optimizer with a learning rate of 0.5 to optimize structural latent features, and the number of iterations is set to 5. Training is performed on a single NVIDIA GeForce RTX 4090 GPU for 1000 epochs.

Table 1: **Quantitative comparisons:** The 3D-CGAN model (Zheng et al., 2023) conditions on material property vectors to synthesize 3D microstructures, while Txt2Microstruct-Net (Zheng et al., 2024) utilizes textual prompts as input to directly generate 3D microstructures. In contrast, we use input-guided generation based on both text and image modalities. **The relatively lower FID score achieved by 3D-CGAN can be attributed to the fact that it is trained on a single class of Voronoi-based open-cell foams. The 3D-CGAN model is only used to compare different inputs.** The first four evaluation metrics are computed over 200 prompts sampled using different random seeds, whereas the coefficient of determination (R^2) is used to assess the goodness of fit between the properties of 2,000 real structures and the predicted properties of the corresponding generated structures. It should be noted that, due to differences in input modalities and training dataset scales, the comparison is not strictly fair.

Method	Accuracy ↑	FID ↓	CLIP ↑	CD ↓	R^2-square ↑	Input
3D-CGAN	–	**53.91**	–	–	–	Vector
Txt2Microstruct-Net	0.8200	77.80	0.5238	0.1162	0.755 **0.757** 0.757	Text
Ours	**0.9150**	**65.55**	**0.5463**	**0.1352**	**0.907 0.713 0.817**	Text, Image



(a) Volume score comparison (b) Young's modulus comparison (c) Anisotropy comparison

Figure 5: **Comparison of target structure and generated structure properties:** For three different properties: (a) volume fraction, (b) Young's modulus, and (c) anisotropy score, we present linear regression plots for 2000 microstructures. A larger R^2 indicates strong linear consistency between the physical and target properties.

4.2 Results of Text-to-Microstructure

We perform qualitative and quantitative evaluations on the dataset to assess the ability of model to generate microstructures under different text and sketch conditions.

Evaluation on Geometries 2000: Table 1 shows that our method achieves a significant improvement in accuracy based on the **Geometries 2000** dataset, indicating stronger consistency between the generated structure and the target material type. The lower FID value also suggests that its distribution is closer to the real samples. As shown in Figure 5, even without explicit attribute conditions, the generated structure still reflects the target physical properties well; except for Young's modulus, all other attributes show good fitting ability. Furthermore, the CLIP score verifies the high semantic consistency between the text description and the generated structure. The CD metric shows that the model fits well on the training set, and although it improves on the test set, it still maintains reasonable geometric similarity.

Evaluation on GenText-Microstruct: We evaluate the model on the GenText-Microstruct dataset, as shown in Table 2. MLLM consistently improves FID and CD over the text-only setting, showing stable gains from multimodal modeling. The text-only variant is limited by the lack of spatial constraints, leading to weaker structural consistency. This does not imply LLM is ineffective, but that it mainly contributes semantic and cross-modal alignment. With additional sketch guidance, structural constraints are strengthened, further improving generation quality and demonstrating complementarity.

Table 2: **Quantitative results on GenText-Microstruct:** The baseline follows a diffusion framework (Yang et al., 2024) with text input, compared against MLLM-based text-to-microstructure methods and their multimodal variants.

Method	FID ↓	CLIP ↑	CD ↓
Guided	84.46	0.3281	0.0666
Ours	**55.47**	**0.4200**	**0.0589**
w/o Sketch	114.17	0.3856	0.1355

Table 3: **Ablation analysis:** We conduct ablation experiments by progressively adding conditions to the text-to-structure LLM generation as a baseline to analyze the impact of different conditions on the results.

Method	Acc. ↑	FID ↓	CLIP ↑
Baseline	0.8600	97.68	0.5269
w/ Properties	0.8950	99.74	0.5242
w/ Sketch	**0.9150**	**65.55**	**0.5463**

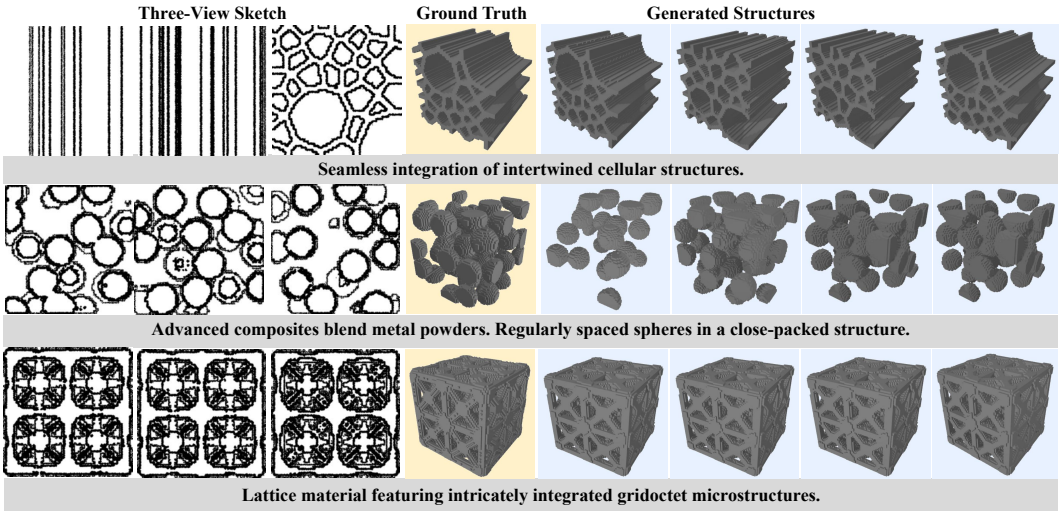


Figure 6: **Qualitative visualizations:** Voxel-based microstructures generated by our model using textual and visual prompts.

Visualizations: Figure 6 illustrates the voxel-based microstructures generated by our model under different text and sketch conditions. The results demonstrate that natural language descriptions effectively guide the generation of structures that align with professional terminology and prescribed geometric characteristics. Moreover, the incorporation of sketch-based conditions enables the model to capture and interpret visual features of the target materials. Overall, the generated microstructures closely resemble, but are not entirely identical to, the real material structures, indicating their strong manufacturability and practical feasibility.

Qualitative Analysis: As shown in Figure 7, we analyzed the impact of attribute constraints on the generated structures through visual ablation experiments. Target attributes were introduced during the generation process to guide the physical properties of the structures, and three representative samples were selected for optimization under three different attributes. The results show that different attribute constraints lead to changes in the voxel distribution and spatial organization of the structures. Although visual differences are not always significant, the overall density and distribution trends are basically consistent with the target attributes, indicating that attribute constraints have statistical effectiveness. Meanwhile, in the process of improving attribute consistency, the structural details and smoothness of some samples decreased, reflecting the trade-off between attribute control and generation fidelity.

4.3 Ablation Studies and Analysis

Ablation Studies: As shown in Table 3, to evaluate the impact of different conditional information, we conduct two text-based ablation studies under identical training settings, focusing on properties and sketch conditions. The results show that introducing property conditions leads to a significant

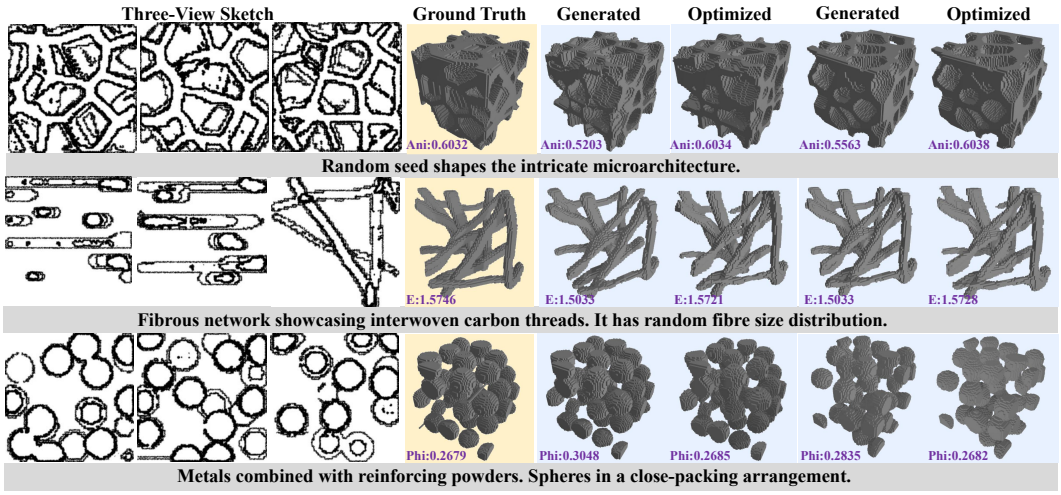


Figure 7: **Visualization of optimization results:** Comparison of generated and optimized structures under three different text and sketch conditions corresponding to anisotropy score, Young's modulus and volume fraction.

improvement in classification accuracy, indicating that textual property information effectively enhances the model's understanding of high-level semantics and category-discriminative features. However, no noticeable improvements are observed in FID and CLIP scores, suggesting that relying solely on textual property information provides limited benefits for reconstructing fine-grained 3D geometric details. Furthermore, we incorporate multi-view sketches into the textual conditions to provide richer spatial structural priors. Compared to above, this approach achieves improvements across all evaluation metrics. In particular, accuracy and FID show more substantial gains, while CLIP also exhibits moderate improvement. These results indicate that multi-view information helps the model better capture 3D structural consistency, although the enhancement in geometric detail remains relatively limited.

Hyperparameter Sensitivity Analysis: We perform a sensitivity analysis of LoRA hyperparameters by varying the r and α. As shown in Table 4, increasing both r and α leads to consistent improvements in Accuracy and CLIP score, while reducing FID and CD. The performance gains are most significant when moving from (64,64) to (128,128), with further increases to (256,256) yielding smaller but noticeable improvements, indicating that the model is moderately sensitive to these hyperparameters.

Table 4: **Hyperparameter sensitivity analysis.** Effect of LoRA rank (r) and scaling factor (α) on model performance.

(r, α)	Acc. ↑	FID ↓	CLIP ↑	CD ↓
(64, 64)	0.7250	527.44	0.4238	0.1515
(128, 128)	0.8100	242.65	0.4527	0.1431
(256, 256)	**0.8600**	**97.68**	**0.5269**	**0.1335**

5 Conclusion

This work tackles text-to-microstructure generation, a key challenge in inverse design. We propose Text-to-Microstructure Large Language Model (T2M-LLM) that combines natural language descriptions with hand-drawn sketches at the visual-level. Specifically, text provides high-level semantic and functional descriptions, while sketches convey low-level geometric and spatial structural information. Furthermore, we incorporate property constraints to optimize the structure to better conform to specific properties. Since benchmarks of text-to-microstructure are limited, we extend existing datasets, construct relevant benchmarks, and implement important baseline approaches. Extensive experiments demonstrate that our method can generate semantically consistent high-quality structures, achieving significant improvements over existing methods. This method has the potential to improve the efficiency and accessibility of microstructure design, while also requiring careful consideration of the reliability of generated structures and the potential misuse of generative models.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and assistance with code development. All AI-assisted content, mathematical derivations, experimental implementations, and generated text were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete mathematical formulation, optimization procedure, and implementation details required to reproduce the proposed method. Hyperparameter settings, evaluation protocols, and additional experimental results are provided in the main paper and appendix.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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.

JB Berger, HNG Wadley, and RM McMeeking. Mechanical metamaterials at the theoretical limit of isotropic elastic stiffness. *Nature*, 543(7646):533–537, 2017.

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.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Collaborative multi-modal coding for high-quality 3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jianpeng Chen, Wangzhi Zhan, Haohui Wang, Zian Jia, Jingru Gan, Junkai Zhang, Jingyuan Qi, Tingwei Chen, Lifu Huang, Muhao Chen, et al. Metamatbench: Integrating heterogeneous data, computational tools, and visual interface for metamaterial discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5334–5344, 2025a.

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.

Yixiao Chen, Zhiyuan Ma, Guoli Jia, Che Jiang, Jianjun Li, and Bowen Zhou. Context-aware autoregressive models for multi-conditional image generation. *arXiv preprint arXiv:2505.12274*, 2025b.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4456–4465, 2023.

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.

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*, pp. 4171–4186, 2019.

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.

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.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4754–4763, 2024.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

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

Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Scalable and composable multimodal text-to-image diffusion models. *arXiv preprint arXiv:2305.15194*, 2023.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506, 2023.

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.

Kang-Hyun Lee and Faez Ahmed. Microlad: 2d-to-3d microstructure reconstruction and generation via latent diffusion and score distillation. *Computer Methods in Applied Mechanics and Engineering*, 449:118527, 2026.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6173–6188, 2024.

Chunping Ma, Yilong Chang, Shuai Wu, and Ruike Renee Zhao. Deep learning-accelerated designs of tunable magneto-mechanical metamaterials. *ACS Applied Materials & Interfaces*, 14(29): 33892–33902, 2022.

Yuki Noguchi, Kei Matsushima, and Takayuki Yamada. Level set-based topology optimization for the design of labyrinthine acoustic metamaterials. *Materials & Design*, 219:110832, 2022.

Johan Phan, Muhammad Sarmad, Leonardo Ruspini, Gabriel Kiss, and Frank Lindseth. Generating 3d images of material microstructures from a single 2d image: a denoising diffusion approach. *Scientific Reports*, 14(1):6498, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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.

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.

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.

Jie Tian, Martin Taylor Sobczak, Dhanush Patil, Jixin Hou, Lin Pang, Arunachalam Ramanathan, Libin Yang, Xianyan Chen, Yuval Golan, Xiaoming Zhai, et al. Crossmatagent: Ai-assisted design of manufacturable metamaterial patterns via multi-agent generative framework. *Advanced Intelligent Discovery*, pp. 202500063, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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*, pp. 6450–6459, 2018.

Than V Tran, SS Nanthakumar, and Xiaoying Zhuang. Deep learning-based framework for the on-demand inverse design of metamaterials with arbitrary target band gap. *npj Artificial Intelligence*, 1(1):2, 2025.

Wenzheng Wu, Kang Wu, Siyuan Xing, Yifei Li, Xiao-Ming Fu, Falai Chen, and Ligang Liu. Ai-driven generation of 3d cad models: A survey. *Computational Visual Media*, 2026.

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.

Mi Xiao, Wei Sha, Yan Zhang, Xiliang Liu, Peigen Li, and Liang Gao. Cmto: Configurable-design-element multiscale topology optimization. *Additive Manufacturing*, 69:103545, 2023.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

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.

Wangzhi Zhan, Jianpeng Chen, Dongqi Fu, and Dawei Zhou. Unimate: A unified model for mechanical metamaterial generation, property prediction, and condition confirmation. *arXiv preprint arXiv:2506.15722*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Chunfeng Zhao, Tian Zhang, Ao Huang, Qiaoyun Wu, and Zhenyu Chen. Deep learning based inverse design of metamaterials for vibration control. *European Journal of Mechanics-A/Solids*, pp. 105957, 2025.

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, 2023a.

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, 2023b.

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.

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.

In this appendix, we provide additional materials to complement the main text. Specifically, we include the construction process of the two datasets (Appendix A.1), the definition of metrics (Appendix A.2) and other qualitative results (Appendix B). These materials aim to facilitate a deeper understanding of our methodology and to support reproducibility of the experiments.

A Experimental Settings

A.1 Dataset Construction

Geometries 2000 covers four major material categories, including metals and alloys, polymers and composites, ceramics, and architected materials. Each sample is associated with aligned text descriptions, material properties, three orthographic views, and the corresponding 3D microstructure. To further enrich the conditioning modalities, we employ SSGA to automatically transform the three-view projections into sketch-based structural priors, resulting in data tuples composed of text, properties, sketches, and 3D microstructures, shown in Figure 8.

In addition, we use the **GenText-Microstructure** dataset, which consists of paired text–structure data where textual descriptions are automatically generated from 3D voxelized structures following (Yang et al., 2024). For each structure, we also derive three-view projections through an automated pipeline and convert them into sketches using SSGA. This dataset contains over 14,000 metamaterial structures that span a wide range of elastic moduli and Poisson's ratios. For every binary voxel structure, mechanical properties are computed from the corresponding stiffness tensor, including C_{11}, C_{12}, C_{44} and V. All numerical properties are normalized using a pretrained scaler. To obtain property-aware textual descriptions, we first employ GPT to generate property-conditioned narratives, which are then refined through a rule-based text generation module that converts quantitative mechanical properties into coherent, scientifically styled natural language that covers bulk stiffness, axial stiffness, Poisson's ratio, and shear strength. Representative samples of this dataset are shown in Figure 9.

A.2 Evaluation Metrics

We assess the performance of the generated microstructures through a set of diverse quantitative metrics designed to capture different aspects of generation quality. Specifically, classification accuracy reflects semantic correctness, FID measures the similarity between generated and real data distributions, the CLIP score evaluates cross-modal consistency between text and structure, and CD characterizes geometric similarity at the shape level. Moreover, the coefficient of determination R^2 is adopted to examine how well the predicted properties of synthesized structures correlate with those of the reference samples. A detailed description of each metric is provided in the following sections.

Classification Accuracy: To evaluate the ability of the **Geometries 2000** dataset to generate category-specific microstructures, we train a 3D convolutional neural network-based classifier on this dataset to classify 20 categories. Higher classification accuracy indicates higher accuracy in generating structures.

CLIP Score: We train a CLIP-style dual encoder on two datasets: a BERT (Devlin et al., 2019) pretrained model for the text encoder and a ResNet3D (Tran et al., 2018) pretrained model for the visual encoder. The CLIP score, calculated by the cosine similarity between the text feature embeddings and the structural feature embeddings, measures the semantic consistency between the text description and the generated structure. A higher score indicates stronger semantic meaning of the structure.

FID Score: We calculate the FID score using the penultimate layer output features of the visual encoder in a CLIP-style dual encoder, with the same training configuration as described above:

$$FID = ||\mu_r - \mu_g||^2 + Tr(C_r + C_g - 2(C_r C_g)^{1/2}), \tag{9}$$

where r and g represent the feature vectors of the real and generated structures, respectively, μ_r and μ_g are the mean vectors, C_r and C_g are the covariance matrices, and Tr represents the trace of the matrix. A lower FID value indicates that the distribution of the generated structure is closer to that of the real samples, suggesting higher generation quality.

CD Score: To quantify geometric similarity, we adopt Chamfer Distance (CD), which measures the discrepancy between point clouds sampled from generated and reference structures. In practice, both

structures are first converted into surface point representations, typically by sampling coordinates from boundary voxels or extracting isosurfaces. Let $P = \{p_i\}_{i=1}^{N}$ and $Q = \{q_j\}_{j=1}^{M}$ represent the point sets sampled from the generated and real structures, respectively. The calculation form of CD is as follows:

$$CD = \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} \|p - q\|_2^2 + \frac{1}{|Q|} \sum_{q \in Q} \min_{p \in P} \|q - p\|_2^2. \tag{10}$$

This metric measures the distance relationship between the two point sets through bidirectional nearest neighbor matching: on the one hand, it measures the degree of fit between the surface points of the generated structure and the surface of the real structure; on the other hand, it evaluates the coverage of the target geometry by the generated result. Therefore, the smaller the CD value, the more consistent the generated structure is with the real structure in terms of geometry.

Coefficient of Determination: To measure the ability of the generated microstructures to maintain consistency with the target values at the physical property level, we train a performance prediction network on **Geometries 2000** that is consistent with the classifier in terms of convolutional structure as a property evaluator. The R^2 coefficient is calculated by comparing the correlation between the generated and real structures in terms of predicted physical properties. A higher R^2 value (close to 1) indicates better fidelity and consistency in physical performance of the generated results.

Inference Efficiency: We supplement our results under the same hardware environment (NVIDIA GeForce RTX 4090). our model's average inference time is approximately 7.3 seconds per sample, while the baseline method (VAE+MLP) is approximately 2.9 seconds per sample. The slower inference of our method is mainly due to the LLM-based autoregressive generation, which generates 512 discrete tokens sequentially, with each step depending on the previous context.

B Qualitative Analysis

As shown in Figure 10, we present more visual ablation experiment results to illustrate the changes brought about by property optimization. To evaluate the impact of target property constraints on the morphology of 3D structures, we present the optimization results of three generated samples for each physical property: Young's modulus, anisotropy, and volume fraction. The visualization shows that the structures generated under different property constraints exhibit systematic differences in voxel distribution and spatial organization. Specifically, the Young's modulus constraint guides changes in the thickness of the structural skeleton and the distribution of local supports; the anisotropy constraint adjusts the structural orientation and uniformity of arrangement; and the volume fraction constraint affects the overall infill density and porosity distribution. Although these changes are not always visually significant, the overall voxel distribution trend remains consistent with the target property, indicating that property constraints effectively regulate structural morphology in a statistical sense. Meanwhile, some samples show a slight decrease in detail and surface smoothness during property enhancement, reflecting the trade-off between property consistency and generation fidelity.

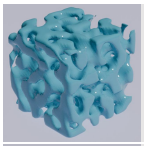
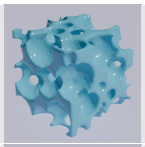
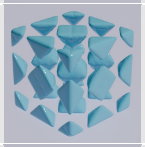
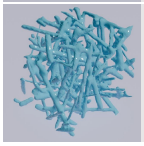
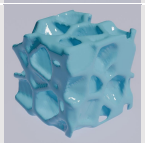
	Metallic matrix with micvoids. Spinodal decomposition leads to intricate patterns. It is a isotropic spinodoid microstructure. E:9.9688, Anisotropy:0.6346, Phi:0.5138		Spherical pores dispersed throughout the foam network, characterized by their diverse sizes influenced by random grain packing. E:6.5287, Anisotropy:0.5983, Phi:0.4046
	Complex matrix showcasing seamless integration of intertwined cellulose threads. It has random fibre size distribution. E:1.5685, Anisotropy:0.7406, Phi:0.0878		Intricate polymer-nanoparticle elements seamlessly integrated. It has same sphere size and random distribution of spheres. E:1.1206, Anisotropy:0.9935, Phi:0.04651
	Powder-enhanced composites with metal matrix. Spheres non-uniformly spaced in a disordered structure. E:1.7615, Anisotropy:0.8708, Phi:0.2073		Mathematically controlled lattice structure from Primitive minimal surface. E:1.9099, Anisotropy:0.6357, Phi:0.2097
	A microstructure composed of linear, cylindrical elements intricately integrated into the structure. E:2.1489, Anisotropy:0.7581, Phi:0.1419		Diverse grain sizes create an intricately foam structure with a constant cell shell thickness. E:9.6627, Anisotropy:0.6328, Phi:0.4379

Figure 8: **Sample Visualization of Geometries 2000:** We present the text, properties, and corresponding 3D microstructures of some samples in this dataset.

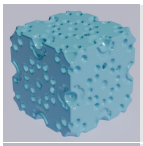
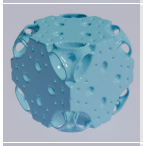
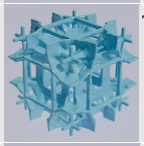
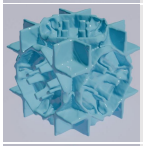
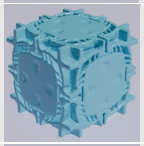
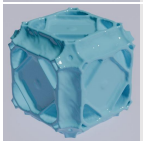
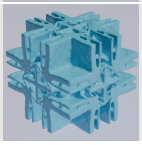
	This structure has containing internal voids, dense solid structure, moderate bulk stiffness, conventional lateral contraction. Phi:0.7208, Poisson:0.3638		This structure has highly compliant in loading, lightweight porous structure, without enclosed voids, positive Poisson's ratio. Phi:0.1802, Poisson:0.4575
	This structure has conventional lateral contraction, moderate shear stiffness, high volume fraction. Phi:0.7467, Poisson:0.3688		This structure has highly compliant in loading, lightweight porous structure, with open connected interior, negative Poisson's ratio. Phi:0.1473, Poisson:-0.1615
	This structure has low axial stiffness, low shear resistance, with enclosed internal cavities, auxetic mechanical response. Phi:0.3127, Poisson:-0.3075		This structure has moderate volume fraction, negative Poisson's ratio, shear-compliant response, compressible under pressure. Phi:0.6494, Poisson:-0.2937
	This structure has without enclosed voids, lightweight porous structure, conventional lateral contraction, low shear resistance. Phi:0.2511, Poisson:0.4084		This structure has compressible under pressure, balanced material density, shear-compliant response. Phi:0.4377, Poisson:-0.2968

Figure 9: **Sample Visualization of GenText-Microstruct:** We also present the text, properties, and corresponding 3D microstructures of some samples in this dataset. It can be observed that the samples in this dataset are different from those in the previous dataset, demonstrating the diversity of the data.

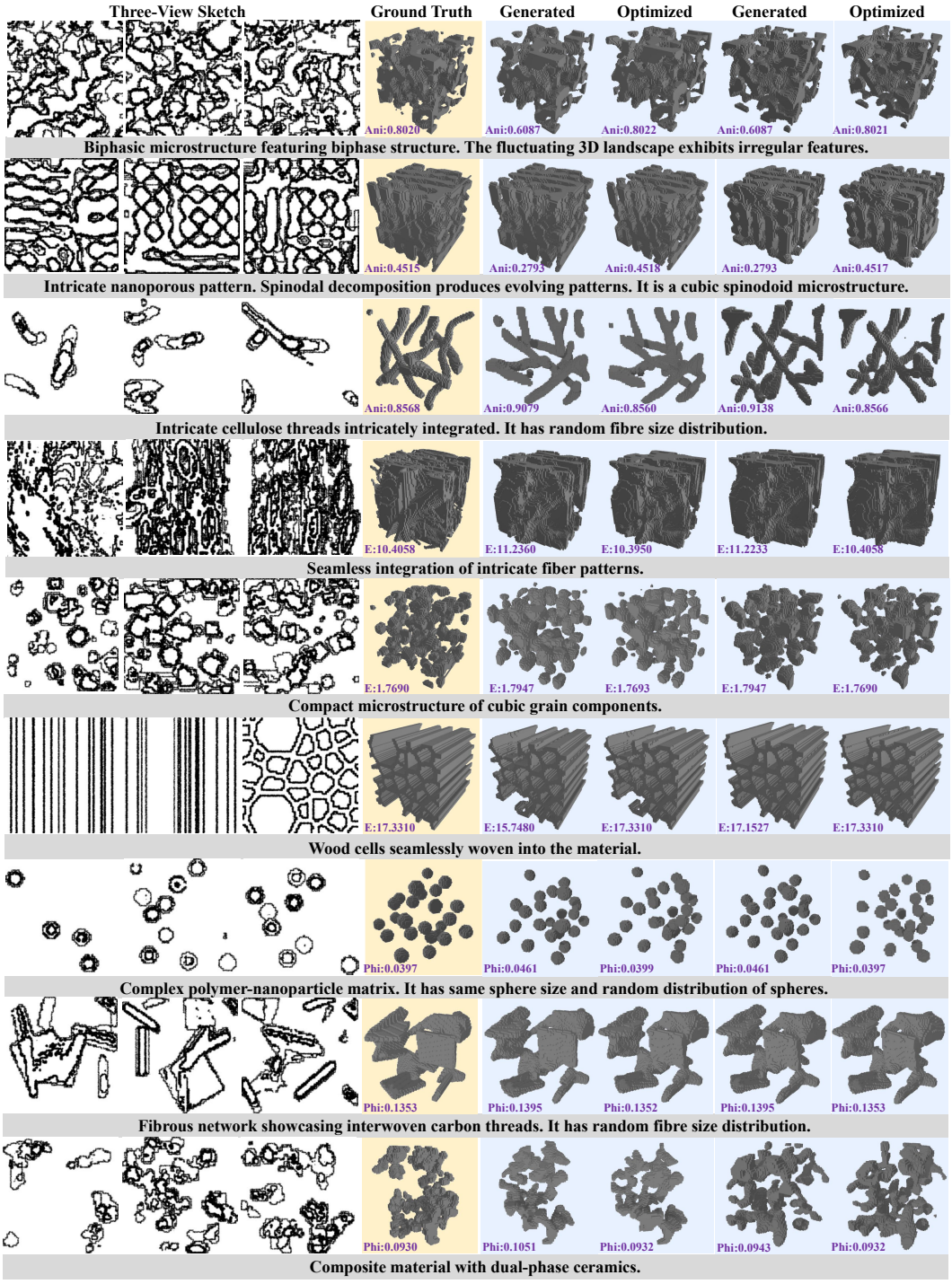


Figure 10: **More visualization of optimization results:** Each property displays a comparison between the generated and optimized structures from three different samples.