# ADAPTIVE LLM-DIFFUSION INTEGRATION FOR ROBUST CRYSTAL GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Text-guided generative models have shown strong promise for crystal generation by incorporating natural-language metadata into the generative process to produce valid and stable materials conditioned on textual descriptions. However, they have some inherent limitations: their LLM backbones are sensitive to the ordering of conditions within the prompt, the LLM and generative model are coupled through a fixed, sample-agnostic diffusion timestep $\tau$, and their effectiveness has largely been demonstrated using only a limited set of LLM architectures. In this work, we introduce AdaCrysGen, a robust and adaptive LLM-diffusion framework for crystal generation. First, we improve robustness to prompt ordering through order-invariant training augmentation and auxiliary condition-specific objectives that encourage consistent generation across different permutations of the conditioning information. Second, we replace the fixed diffusion timestep with a lightweight per-sample policy network $\pi_\phi$ that adaptively determines the amount of diffusion refinement required for each LLM-generated structure. Finally, we benchmark AdaCrysGen across a diverse collection of open-source LLMs spanning multiple model families and scales, and observe that AdaCrysGen, across all backbones, consistently outperforms existing baselines in generating valid, novel, and stable materials, providing a broad assessment of the generality and robustness of LLM-diffusion integration for crystal generation.

## 1 INTRODUCTION

The discovery of 3D atomic structures of novel crystal materials is vital for developing next-generation functional materials such as batteries, semiconductors, solar cells, electrolytes, and catalysts (Butler et al., 2018; Desiraju, 2002). However, traditional approaches such as density functional theory (DFT) (Jones & Gunnarsson, 1989; Parr & Weitao, 1995; Sholl & Steckel, 2009) are computationally expensive, making them impractical for exhaustive large-scale exploration. In recent years, generative models, including diffusion (Xie et al., 2021; Luo et al., 2023; Jiao et al., 2023; Zeni et al., 2023) and flow-based models (Miller et al., 2024), as well as large language models (LLMs) (Antunes et al., 2024; Gruver et al., 2024), have gained significant traction as promising alternatives, offering both effectiveness and efficiency. These models learn the distribution of stable crystal structures from data under various conditional or unconditional generation settings. Lately, text-guided generative models have been proposed that incorporate natural-language metadata into crystal generation by using an LLM as part of the generative process. Models such as CrysLLMGen (Khastagir et al., 2026) and FlowLLM (Sriram et al., 2024) combine language-based priors with diffusion- and flow-based generators, respectively. In these models, first, the LLM generates the initial discrete atom types together with continuous atomic coordinates and lattice structure, while the subsequent diffusion- or flow-based model further refines the atomic coordinates and lattice structure.

Although these hybrid models improve the validity, novelty, and stability of generated crystal materials, they still have some inherent limitations. **First,** existing hybrid frameworks, such as CrysLLMGen or FlowLLM, have only been validated on a single, comparatively small language model (LLaMA-2) as the backbone. It therefore remains unclear whether the benefits of LLM-diffusion coupling generalize to more capable and diverse LLM backbones, as systematic benchmarking across stronger and more varied architectures is still missing. **Second,** existing models finetune LLM backbones using the conventional causal language modeling objective (next-token prediction) on prompts that present conditioning information in a fixed order. As a result, these models can learn to associate a condition's meaning with its position in the prompt rather than its content, and shuffling the condition order at



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.

inference time can significantly degrade generation quality. In particular, we observe a substantial drop in the model's ability to generate the correct atomic composition and chemical formula (Fig. 2) specified in the prompt. Composition prediction is especially important in hybrid models because the subsequent denoising stage only refines the atomic coordinates and lattice structure, while the atom types predicted by the LLM remain fixed. Therefore, any composition error the LLM introduces under prompt shuffling cannot be corrected downstream during the denoising stage. **Finally,** while structural errors introduced by the LLM may still be correctable by the diffusion stage, existing hybrid models do not fully exploit this. In frameworks such as CrysLLMGen, the LLM and diffusion model are trained independently and coupled only at sampling time through a fixed hyperparameter $\tau$, selected using validation data, which determines the denoising step of the diffusion model at which the LLM-generated intermediate structure is passed to for further refinement. Using the same $\tau$ for every sample ignores how much geometric correction each structure actually requires, leading to suboptimal generation quality.

In this work, we address these limitations by introducing **AdaCrysGen**, an ***Ada***ptive LLM-Diffusion framework for robust ***Crys***tal ***Gen***eration, which follows a hybrid design with two components: a fine-tuned LLM and a pre-trained diffusion model. First, to reduce the LLM's sensitivity to shuffled prompt ordering, we introduce an order-invariant training augmentation, where the conditioning sentences in each training prompt are randomly reordered during fine-tuning. This encourages the model to identify and use each condition based on its content rather than its position. Additionally, we introduce an auxiliary loss for each condition, added to the standard LLM objective, to encourage the LLM to generate crystal structures aligned with the conditions specified in the prompt. Next, we train a lightweight policy network that predicts a per-sample diffusion injection timestep $\tau^*$ from the LLM's own output, replacing CrysLLMGen's single fixed $\tau$ with an adaptive one. We keep both the LLM and diffusion model frozen, and only this small coupling mechanism is learned. The policy is trained offline on a reward derived from post-relaxation thermodynamic stability, allowing each sample to receive exactly as much diffusion-based structural correction as it needs. Finally, we empirically validate AdaCrysGen by benchmarking the hybrid framework across 14 LLM backbones spanning multiple model families and scales, including LLaMA-2, LLaMA-3, Qwen3, and Gemma3. Extensive experiments on benchmark dataset (MP-20) show that AdaCrysGen, across all backbones, consistently outperforms existing baselines in generating valid, novel, and stable materials (in SUN metric, improving by over $\mathbf{99.4}\%$ and $\mathbf{329.1}\%$ over the strongest diffusion and flow-matching baselines, and $\mathbf{120.7}\%$ and $\mathbf{16.0}\%$ over the matched LLaMA-2-7B LLM-only and fixed-$\tau$ hybrid baselines.) Notably, each hybrid variant outperforms its corresponding standalone LLM, demonstrating that the benefit of LLM–diffusion coupling generalizes across architectures. Moreover, we find that the LLM becomes substantially more robust to prompt condition ordering after fine-tuning, indicating that robustness to prompt ordering and structural quality benefit from complementary mechanisms on the LLM and diffusion sides, respectively. To sum up, our novel contributions in this work are as follows:

- **Adaptive LLM-Diffusion Integration:** We propose AdaCrysGen, a hybrid LLM-diffusion framework for robust crystal generation, where we replace the fixed diffusion injection timestep $\tau$ used in prior work with a per-sample adaptive $\tau^*$, improving structural validity and stability of generated materials.
- **Robustness to Prompt Condition Ordering:** We introduce an order-invariant training augmentation and auxiliary condition-specific losses that make LLM-based generation robust to the ordering of conditions in the input prompt, and show that the resulting model better satisfies the specified conditions regardless of prompt ordering.
- **Multi-Backbone Generalization:** We benchmark the hybrid LLM-diffusion framework across multiple open-source LLM backbones spanning different families and scales, and show it consistently outperforms the corresponding standalone LLM in every case, demonstrating that the benefit of LLM-diffusion coupling generalizes across architectures.

## 2 PRELIMINARIES

### 2.1 PROBLEM FORMULATIONS

We study the task of generating novel 3D crystal structures with the objective of discovering valid and stable materials. Let $\mathcal{C} = \{\mathbf{C}_i\}_{i=1}^{K}$ denote a dataset of known valid and stable material structures sampled from an unknown underlying crystal distribution $p(\mathbf{C})$. Our goal is to develop a generative

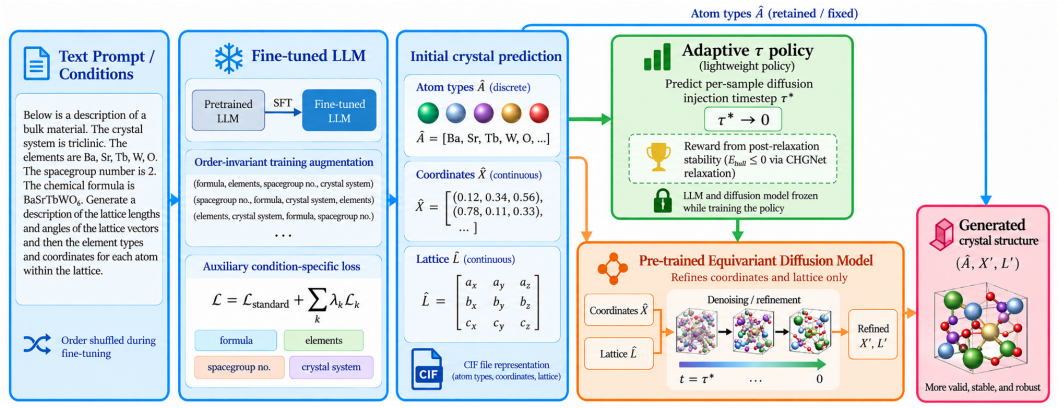



Figure 1: Model architecture of AdaCrysGen. Given a text prompt, the fine-tuned LLM generates an initial prediction for $\hat{\mathbf{A}}$, $\hat{\mathbf{X}}$, and $\hat{\mathbf{L}}$. $\hat{\mathbf{A}}$ is retained as the final composition, while a lightweight policy $\pi_\phi$ predicts a per-sample adaptive timestep $\tau^*$ from the LLM's output; the $\hat{\mathbf{X}}$ and $\hat{\mathbf{L}}$ are then refined by a pre-trained diffusion model, injected at $\tau^*$, to produce the final generated crystal structure.

model that learns a parameterized distribution $f_\theta(\mathbf{C})$ that approximates $p(\mathbf{C})$. A key requirement for crystal generative models is to ensure that the learned distribution $f_\theta(\mathbf{C})$ respects periodic E(3) invariance (described in 2.2), such that equivalent crystal representations are treated consistently.

### 2.2 Crystal Structure Representation and Symmetries

A crystal can be represented as a periodic 3D arrangement of atoms, where a unit cell is repeatedly translated along three lattice directions to construct the complete structure. For a unit cell containing $N$ atoms, we represent the crystal using three components: the atom types, atomic coordinates, and lattice vectors. The atom identities are encoded by the atom type matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N]^T \in \mathbb{R}^{N \times k}$, where each $a_i$ is a one-hot representation of the atom type and $k$ denotes the number of possible atomic species. The atomic positions are represented by the Coordinate Matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times 3}$, where $x_i \in \mathbb{R}^3$ gives the position of the $i^{th}$ atom within the unit cell. The periodicity of the crystal is defined by the Lattice Matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]^T \in \mathbb{R}^{3 \times 3}$, whose vectors determine the three directions along which the unit cell is repeated.

A valid crystal representation should be invariant under permutation, translation, rotation, and periodic transformations (Dresselhaus et al., 2007; Zee, 2016). In particular, reordering atoms or applying a global translation to all atomic positions should not change the represented material. Similarly, rotating the atomic coordinates and lattice together should preserve the crystal representation. Moreover, periodic invariance further reflects the repeating nature of crystals, where different but equivalent choices of unit cells and coordinates can describe the same underlying material. Hence, the representation $(\mathbf{A}, \mathbf{X}, \mathbf{L})$ is not unique: a physically identical crystal admits multiple such representations related by permutation, translation, rotation, and periodic invariance. More details about these invariance properties are provided in Appendix C.2.

## 3 Proposed Methodology: AdaCrysGen

We introduce AdaCrysGen, a hybrid LLM-diffusion framework (Figure 1) consisting of two components: a large language model (LLM) and a diffusion model. This design is motivated by prior work (Khastagir et al., 2026; Sriram et al., 2024), which shows that LLMs are effective at predicting discrete atom types and achieving high compositional validity, whereas denoising-based models are inherently better suited to continuous variables and preserving geometric equivariance, resulting in higher structural validity and stability. Hybrid models aim to combine these complementary strengths, mitigating the limitations of each approach individually. In AdaCrysGen, we fine-tune a pretrained LLM and train a diffusion model independently on the same training dataset. During sampling, the LLM first generates initial predictions of atom types ($\hat{\mathbf{A}}$), coordinates ($\hat{\mathbf{X}}$), and lattice ($\hat{\mathbf{L}}$). We retain the predicted atom types and pass $\hat{\mathbf{X}}$, $\hat{\mathbf{L}}$ to the diffusion model, which refines these continuous components to improve structural validity and stability. Unlike CrysLLMGen, however,

we introduce two key changes. First, during LLM fine-tuning, we apply (i) an order-invariant training augmentation and (ii) a weighted auxiliary loss for each condition specified in the input prompt. Second, for adaptive sampling, we train a lightweight policy network that predicts a per-sample diffusion injection timestep $\tau^*$ directly from the LLM's output, replacing CrysLLMGen's single fixed $\tau$ with an adaptive one. This allows each generated sample to receive exactly as much diffusion-based refinement of its coordinates and lattice as it actually requires, rather than applying a uniform, fixed correction to every sample regardless of how far its predicted geometry is from a valid and stable structure.

### 3.1 Large Language Model (LLM) for Crystal Material Generation

We begin by describing the LLM module of AdaCrysGen and its fine-tuning paradigm. We build upon prior work (Gruver et al., 2024) to develop our base LLM module, in which crystal structures are converted into sequential text representations and fed into the LLM for further processing. We transform a dataset of crystal structures $\mathcal{C} = \{\mathbf{C}_i\}_{i=1}^{K}$ into a corresponding dataset of sequences $\mathcal{W} = \{W_i\}_{i=1}^{K}$, where each sequence $W_i$ represents the CIF text format of the crystal structure $\mathbf{C}_i$. Each sequence $W_i$ is further tokenized using the pretrained LLM's own tokenizer, without introducing any crystal-specific vocabulary. We fine-tune a pretrained LLM on this dataset of crystal structures represented as text sequences, using task-specific prompts to support both unconditional and text-conditional generation. We adopt the prompt formats proposed in (Gruver et al., 2024) for both tasks. Specifically, for conditional generation, each training prompt specifies a random subset of target conditions, such as the chemical formula, elements, space group number, and crystal system that are expressed as natural-language sentences preceding the target crystal sequence.

However, existing prior works fine-tune LLMs using the conventional causal language modeling objective (next-token prediction) on prompts that present conditioning information in a fixed order. As a result, these models can learn to associate a condition's meaning with its position in the prompt rather than its content, relying on positional or template-based shortcuts instead of robustly interpreting the conditions themselves. Consequently, shuffling the order of these conditions at inference time can significantly degrade generation quality (Section 4.3). In particular, we observe a substantial drop in the model's ability to generate the correct atomic composition and chemical formula specified in the prompt. Composition prediction is especially important in hybrid models, since the subsequent denoising stage only refines the atomic coordinates and lattice structure, while the atom types predicted by the LLM remain fixed. Hence, any composition error the LLM introduces under prompt shuffling cannot be corrected downstream during the denoising stage. Additionally, prior hybrid frameworks have only been validated on a single LLM backbone (LLaMA-2), leaving open the question of whether the advantage of LLM-diffusion coupling over a standalone LLM is a genuine, architecture-agnostic property of the coupling itself, or an artifact specific to that one backbone. Taken together, these observations point to a broader lack of robustness and generalizability in the generation capabilities of existing hybrid frameworks.

To address these limitations, we introduce two simple yet effective modifications to the LLM fine-tuning algorithm, aimed at improving the robustness of material generation. First, we apply an *order-invariant training augmentation*: the ordering of the conditioning sentences in each training prompt is randomly permuted, encouraging the model to identify each condition by its content rather than its position. Second, we introduce an *auxiliary condition-specific loss*, added to the standard next-token prediction objective with a tunable weight, that explicitly supervises the tokens most directly responsible for satisfying each condition. Hence, the combined loss for LLM fine-tuning is

$$\mathcal{L}_{\text{LLM}} = \mathcal{L}_{\text{standard}} + \sum_{k \in \mathcal{K}} \lambda_k \, \mathcal{L}_k, \tag{1}$$

where $\mathcal{L}_{\text{standard}}$ is the standard token-averaged cross-entropy loss over the full sequence, $\mathcal{K}$ is the set of conditioning attributes (e.g., chemical formula, elements, space group number, crystal system), $\mathcal{L}_k$ is the auxiliary loss for condition $k$, and $\lambda_k$ is a tunable weight controlling its contribution. Together with the order-invariant augmentation, these auxiliary losses encourage the LLM to generate crystal structures that are more robust to prompt ordering and better aligned with the specified conditions (Section 4.3). We fine-tune the LLM module on the training data using Low-Rank Adaptation (LoRA) (Hu et al., 2021), starting from a pretrained LLM and adapting it to crystal structures represented as text sequences, guided by task-specific prompts for either unconditional or conditional generation. Our fine-tuning procedure is backbone-agnostic: we explore four distinct classes of open-source LLMs, spanning multiple model families and scales, with benchmark results across all

backbones presented in Section 4.1. These results demonstrate that the benefits of LLM-diffusion coupling generalize across different LLM backbones, rather than being specific to any single one.

### 3.2 Diffusion Model for Crystal Structure Refinement

Once the LLM module generates initial predictions for the atomic composition, coordinates, and lattice parameters, we retain the predicted atom types as the final composition of the generated material, while the corresponding atomic coordinates and lattice parameters are further refined by a denoising generative model. Prior work has explored different choices for this refinement stage: CrysLLMGen (Khastagir et al., 2026) couples the LLM with an equivariant diffusion model, while FlowLLM (Sriram et al., 2024) couples it with a flow-matching algorithm. Khastagir et al. (2026) report that their diffusion-based hybrid model achieves higher compositional validity, stability, and property (density) alignment than the flow-matching-based FlowLLM. We therefore adopt a diffusion-based refinement model in our setup: we train the diffusion model on a structure prediction task, where, given the atom types $\mathbf{A}$ of a crystal material, it learns the conditional distribution $p(\mathbf{X}, \mathbf{L} \mid \mathbf{A})$.

**Diffusion on X and L.** Fractional atomic coordinates $\mathbf{X}$ reside in the periodic, bounded quotient space $\mathbb{R}^{N\times 3}/\mathbb{Z}^{N\times 3}$, so the standard Gaussian noise used in DDPM is not well suited to $\mathbf{X}$. Following DiffCSP (Jiao et al., 2023), we instead corrupt $\mathbf{X}$ with a Wrapped Normal distribution (De Bortoli et al., 2022), giving the forward transition $q(\mathbf{X}_t \mid \mathbf{X}_0) = \mathcal{N}_W(\mathbf{X}_t \mid \mathbf{X}_0, \sigma_t^2\mathbf{I})$, or equivalently $\mathbf{X}_t = f_w\left(\mathbf{X}_0 + \sigma_t \boldsymbol{\epsilon}^{\mathbf{X}}\right)$ where $\boldsymbol{\epsilon}^{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\sigma_t$ follows an exponential noise schedule and $f_w(\mathbf{X}) = \mathbf{X} - \lfloor \mathbf{X} \rfloor$ truncates each entry to its fractional part, ensuring periodic translation equivariance. In contrast, the lattice matrix $\mathbf{L}$ is unconstrained and continuous; therefore, we corrupt it using a standard DDPM process (Ho et al., 2020): $\mathbf{L}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{L}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}^{\mathbf{L}}$ with $\boldsymbol{\epsilon}^{\mathbf{L}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$ and $\alpha_t = 1 - \beta_t$ for a noise schedule $\{\beta_t\}_{t=1}^{T}$. A single denoising network $\Phi_\theta(\mathbf{A}, \mathbf{X}_t, \mathbf{L}_t, t)$, built on an equivariant graph neural network (following CSPNet (Jiao et al., 2023)), jointly predicts the coordinate and lattice denoising terms $\hat{\boldsymbol{\epsilon}}^{\mathbf{X}}$ and $\hat{\boldsymbol{\epsilon}}^{\mathbf{L}}$ at each step, using periodic-invariant message passing to respect the symmetries described in Section 2.2. The network is trained independently with a combined score-matching and denoising objective over $\mathbf{X}$ and $\mathbf{L}$, respectively:

$$\mathcal{L}_{\text{refinement}} = \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{lattice}} \tag{2}$$

$$\mathcal{L}_{\text{coord}} = \mathbb{E}_{\mathbf{X}_t \sim q(\mathbf{X}_t \mid \mathbf{X}_0),\, t \sim \mathcal{U}(1,T)} \left\| \nabla_{\mathbf{X}_t} \log q(\mathbf{X}_t \mid \mathbf{X}_0) - \hat{\boldsymbol{\epsilon}}^{\mathbf{X}} \right\|_2^2 \tag{3}$$

$$\mathcal{L}_{\text{lattice}} = \mathbb{E}_{\boldsymbol{\epsilon}^{\mathbf{L}},\, t \sim \mathcal{U}(1,T)} \left\| \boldsymbol{\epsilon}^{\mathbf{L}} - \hat{\boldsymbol{\epsilon}}^{\mathbf{L}} \right\|_2^2 \tag{4}$$

### 3.3 Adaptive Diffusion Integration via Learned $\tau^*$

As described in Section 3.1, existing hybrid frameworks such as CrysLLMGen inject the LLM-generated coordinates $\hat{\mathbf{X}}$ and lattice $\hat{\mathbf{L}}$ into the diffusion model at a single, fixed timestep $\tau$, selected once using validation data, and apply refinement identically to every sample. Such a one-size-fits-all strategy can be suboptimal, since the amount of refinement a given sample actually needs depends on how consistent and well-formed the LLM's own output already is. We therefore replace this fixed $\tau$ with a per-sample adaptive value $\tau_i^*$, predicted directly from the LLM's output by a lightweight policy network $\pi_\phi(\tau \mid \mathbf{c}_i)$, while keeping both the LLM and the diffusion model entirely frozen. This is a deliberate design choice that isolates the coupling mechanism as the sole point of adaptation, rather than reshaping either pretrained generative model directly.

To begin, the context vector $\mathbf{c}_i$ is computed from the LLM's frozen output $(\hat{\mathbf{A}}_i, \hat{\mathbf{X}}_i, \hat{\mathbf{L}}_i)$ before any diffusion refinement takes place, and it summarizes the lattice lengths and angles, the mean and standard deviation of the fractional coordinates, atom-type statistics, and the compositional entropy of $\hat{\mathbf{A}}_i$. The policy network $\pi_\phi$ is a small multi-layer perceptron that outputs a categorical distribution over a discrete set of candidate timesteps $\mathcal{T} = \{\tau_1, \ldots, \tau_B\}$ spanning the diffusion model's timestep range, with the adaptive timestep at inference time given by

$$\tau_i^* = \arg\max_{\tau \in \mathcal{T}} \pi_\phi(\tau \mid \mathbf{c}_i). \tag{5}$$

Since evaluating a candidate $\tau$ requires running the (frozen) diffusion model and scoring the resulting structure, we precompute this signal offline over the validation set. For every sample $i$ and every $\tau \in \mathcal{T}$, we inject $(\hat{\mathbf{X}}_i, \hat{\mathbf{L}}_i)$ into the diffusion model at $\tau$, relax the resulting structure with CHGNet, and compute a binary stability reward against a precomputed Materials Project phase diagram:

$$R(i, \tau) = \mathbb{1}\left[E_{\text{hull}}(i, \tau) \leq 0.0 \text{ eV/atom}\right], \tag{6}$$

| Models | Validity(%) ↑ | | Property Statistics↓ | | Stability, Uniqueness and Novelty(%) ↑ | | | | Avg. Rank† |
|---|---|---|---|---|---|---|---|---|---|
| | Compositional | Structural | Density($d_\rho$) | Element($d_{\text{elem}}$) | Meta-stable | MSUN | Stable | SUN | |
| CDVAE | 86.70 | **100** | 0.687 | 1.432 | 23.58 | 21.99 | 3.08 | 2.56 | 13.13 |
| DiffCSP | 83.25 | **100** | 0.352 | 0.339 | 35.04 | 32.19 | 7.36 | 5.61 | 10.63 |
| DiffCSP++ | 85.12 | <u>99.94</u> | 0.235 | 0.375 | 42.39 | 30.56 | 8.58 | 6.55 | 10.81 |
| MatterGen | 86.34 | **100** | 0.459 | 0.254 | 78.00 | 21.00 | 13.00 | 7.10 | 9.00 |
| SymmCD | 87.13 | 92.30 | 0.531 | 0.210 | 40.01 | 31.69 | 9.99 | 6.76 | 10.25 |
| ADiT | 90.76 | 99.34 | **0.076** | <u>0.056</u> | 81.60 | 25.91 | 14.10 | 4.70 | 7.75 |
| FlowMM | 83.19 | 96.85 | 0.239 | 0.083 | 31.64 | 22.46 | 4.76 | 3.06 | 12.00 |
| MCFlow | 89.85 | 99.44 | <u>0.136</u> | **0.055** | 80.30 | 22.10 | 15.10 | 3.30 | 8.00 |
| Crystal-Text-LLM (LLaMA-2 7B) | 91.19 | 97.70 | 1.575 | 0.271 | 56.60 | 26.66 | 12.67 | 4.84 | 11.50 |
| FlowLLM (LLaMA-2 7B) | 91.19 | <u>99.94</u> | 1.142 | 0.271 | 60.19 | 30.17 | 14.66 | 6.97 | 9.81 |
| CrysLLMGen (LLaMA-2 7B) | 91.19 | **100** | 1.105 | 0.271 | 62.02 | 35.94 | 16.79 | 9.21 | 7.88 |
| AdaCrysGen (LLaMA-2 7B) | 91.19 | **100** | 1.0655 | 0.271 | 71.80 | 45.68 | 18.64 | 10.68 | 6.88 |
| AdaCrysGen (LLaMA-2 13B) | <u>91.96</u> | **100** | 1.0640 | 0.238 | 75.64 | <u>45.81</u> | 22.08 | 12.43 | 5.38 |
| AdaCrysGen (LLaMA-3 8B) | **93.44** | **100** | 1.0205 | 0.185 | <u>83.78</u> | 44.84 | <u>24.52</u> | 12.64 | <u>4.00</u> |
| AdaCrysGen (Gemma3-12B) | 91.64 | **100** | 1.0672 | 0.192 | **87.92** | **48.27** | **26.36** | **14.16** | **3.75** |
| AdaCrysGen (Qwen3-14B) | 91.12 | **100** | 0.5647 | 0.231 | 78.94 | 45.12 | 22.16 | <u>13.88</u> | 5.25 |

Table 1: Summary of results on *De Novo Generation Task* of different classes of generative models on the MP-20 dataset across validity, property statistics, and stability metrics. **Bold** denotes the best value and <u>underline</u> denotes the second-best value in each column. †Avg. Rank denotes the mean rank across all eight metrics (lower is better).

We use stability as the reward signal since, unlike uniqueness and novelty, it can be evaluated independently for each sample immediately after relaxation, making it well suited to a per-sample, offline-trained policy. This yields a logged dataset $\mathcal{D} = \{(\mathbf{c}_i, \tau, R(i,\tau))\}$ of context-action-reward triples. Because selecting $\tau_i^*$ is a single, one-shot decision per sample rather than part of a sequential trajectory, we train $\pi_\phi$ entirely offline on $\mathcal{D}$, using a *reward-weighted* objective where each sample's rewards are converted into a soft target distribution $p_i(\tau) \propto \exp(R(i,\tau)/\lambda)$ over $\mathcal{T}$, and $\pi_\phi$ is trained to match it via KL divergence,

$$\mathcal{L}_{\text{reward}}(\phi) = \mathbb{E}_i \left[ D_{\text{KL}} \left( p_i(\cdot) \,\|\, \pi_\phi(\cdot \mid \mathbf{c}_i) \right) \right]. \tag{7}$$

At sampling time, we compute $\tau_i^*$ for every sample via a single, cheap forward pass through $\pi_\phi$, group samples by their predicted timestep, and invoke the diffusion model's existing batched sampler once per group. This allows adaptive, per-sample refinement to be introduced without altering either pretrained backbone.

### 3.4 Sampling Procedure for AdaCrysGen

Once trained, the sampling of AdaCrysGen proceeds in three stages. First, the pretrained LLM generates an initial prediction for the crystal material given a text prompt $\mathcal{P}$ (conditional/unconditional), producing the atom types $\hat{\mathbf{A}}$, fractional coordinates $\hat{\mathbf{X}}$, and lattice matrix $\hat{\mathbf{L}}$. The predicted atom types $\hat{\mathbf{A}}$ are retained as the final composition of the generated material and are not modified further at any later stage. Second, rather than passing $\hat{\mathbf{X}}$ and $\hat{\mathbf{L}}$ directly to the diffusion model at a single, fixed timestep, we compute a context vector $\mathbf{c}$ from the LLM's output and use the trained policy network $\pi_\phi$ to select a per-sample adaptive timestep $\tau^*$ from the candidate set $\mathcal{T}$. Third, the predicted coordinates and lattice are refined by the pretrained diffusion model, which injects $(\hat{\mathbf{X}}, \hat{\mathbf{L}})$ at the sample-specific timestep $\tau^*$ and runs the reverse diffusion process from $\tau^*$ down to 0. The refined coordinates and lattice, together with the retained atom types, constitute the final generated crystal material $\mathbf{C}_{\text{new}}$. The detailed sampling algorithm is provided in Algorithm 1 in the Appendix.

## 4 Experiments

### 4.1 De Novo Material Generation (DNG) Task

**Setup:** De Novo Material Generation (DNG) is an unconditional generation task in which the model generates atom types, coordinates and lattice parameters for a valid crystal material, aimed at producing structures that are both novel and thermodynamically stable. Following prior work, we evaluate AdaCrysGen on the DNG task using the widely-used MP-20 dataset (Xie et al., 2021). MP-20 comprises of 45,231 materials from the Materials Project (Jain et al., 2013) database, that

contain at most 20 atoms per unit cell and reside within 0.08 eV/atom of the convex hull energy. We compare AdaCrysGen against popular state-of-the-art generative models across different categories: CDVAE (Xie et al., 2021), DiffCSP (Jiao et al., 2023), DiffCSP++ (Jiao et al., 2024), MatterGen (Zeni et al., 2023), SymmCD (Levy et al., 2025), and ADiT (Joshi et al., 2025) from the class of diffusion models; FlowMM (Miller et al., 2024) and MCFlow (Seong et al., 2026) from the flow matching paradigm; Crystal-Text-LLM (Gruver et al., 2024) representing the LLM-based approach; and FlowLLM (Sriram et al., 2024), CrysLLMGen (Khastagir et al., 2026) and different variants of AdaCrysGen representing hybrid models. While training all competitive models and AdaCrysGen, we followed the standard dataset split of 60% for training, 20% for validation, and 20% for testing.

**Evaluation:** We generate 10,000 materials using each model and assess the quality of the generated materials across three broad categories of popular metrics: (1) Validity Metrics, (2) Property Statistics, following (Xie et al., 2021), and (3) Stability Metrics, following (Miller et al., 2024). For validity, we assess both structural and compositional validity: structural validity requires all pairwise atomic distances within a crystal to exceed $0.5$ Å, while compositional validity is assessed via SMACT (Davies et al., 2019) rules to confirm charge neutrality. For property statistics, we evaluate statistical fidelity by computing $d_\rho$ and $d_{\text{elem}}$, the Wasserstein distance between the generated and test data distributions for density and elemental composition, respectively. Finally, we assess thermodynamic stability by computing the energy above the convex hull ($E_{\text{hull}}$) after structural relaxation with CHGNet (Deng et al., 2023). We identify a crystal as stable if $E_{\text{hull}} \leq 0.0$ eV/atom and as metastable if $E_{\text{hull}} \leq 0.1$ eV/atom, and report Meta-Stable Rate (%) and Stable Rate (%) as the percentage of the 10,000 generated structures satisfying each threshold, respectively. Among the metastable and stable structures, we further report the percentage that are also unique and novel, as MSUN and SUN, respectively. We report all results for the DNG task in Table 1. For AdaCrysGen, we benchmark 14 LLM backbones spanning different model families and scales; Table 1 reports results for 5 representative backbones, with the remaining 9 reported in Appendix (Table 3).

**Results:** Overall, every variant of AdaCrysGen surpasses existing baselines on most metrics, achieving superior (i.e., lower) average rank among all compared models in Table 1. We highlight several key takeaways from the results below.

In terms of stability, all variants of AdaCrysGen excel at generating stable, unique, and novel materials, substantially outperforming standalone denoising-based generative models, including both diffusion and flow-matching approaches. First, compared with the strongest diffusion baseline on each respective stability metric, AdaCrysGen(Gemma3-12B) improves Meta-Stable by $\mathbf{7.7}\%$, MSUN by $\mathbf{50.0}\%$, Stable by $\mathbf{87.0}\%$, and SUN by $\mathbf{99.4}\%$. The gains are particularly pronounced on the stricter MSUN and SUN metrics, with AdaCrysGen nearly doubling the SUN rate of the strongest diffusion-based baseline. A similar trend holds against flow-matching approaches: compared with MCFlow, the Gemma3-12B variant improves Meta-Stable, MSUN, Stable, and SUN by $\mathbf{9.5}\%$, $\mathbf{118.4}\%$, $\mathbf{74.6}\%$, and $\mathbf{329.1}\%$, respectively. The other variants of AdaCrysGen also produce very competitive performance. Next, compared with Crystal-Text-LLM, all variants of AdaCrysGen achieve better performance on all four metrics. Focusing on the matched LLaMA-2 7B backbone, AdaCrysGen (LLaMA-2 7B) improves Meta-Stable, MSUN, Stable, and SUN by $\mathbf{26.9}\%$, $\mathbf{71.3}\%$, $\mathbf{47.1}\%$, and $\mathbf{120.7}\%$, respectively. This highlights the benefit of refining the initial LLM-generated structures, substantially increasing the fraction of generated materials that simultaneously satisfy stability, uniqueness, and novelty. Finally, comparing against hybrid approaches, including FlowLLM and CrysLLMGen, all variants of AdaCrysGen again achieve better performance on all four metrics. In particular, using the same LLaMA-2 7B backbone, AdaCrysGen improves over FlowLLM by $\mathbf{19.3}\%$, $\mathbf{51.4}\%$, $\mathbf{27.1}\%$, and $\mathbf{53.2}\%$ on Meta-Stable, MSUN, Stable, and SUN, respectively. More importantly, compared with CrysLLMGen, which also uses the same LLaMA-2 7B backbone but employs a single, globally selected diffusion timestep for all generated samples, AdaCrysGen improves the corresponding metrics by $\mathbf{15.8}\%$, $\mathbf{27.1}\%$, $\mathbf{11.0}\%$, and $\mathbf{16.0}\%$, respectively. Since the LLM backbone is held fixed in these comparisons, the consistent improvements demonstrate the benefit of sample-adaptive denoising over applying a fixed, sample-agnostic refinement strategy uniformly across generated crystals.

On validity metrics, denoising (diffusion/flow-matching) based models are generally effective at modeling atomic coordinates and lattice parameters, and consequently achieve high structural validity ($\approx 100\%$), while LLMs are good at modeling atomic composition and achieve strong compositional validity ($\geq 90\%$). Hybrid models such as FlowLLM and CrysLLMGen balance this trade-off, achieving strong performance on both fronts simultaneously. AdaCrysGen reinforces this hypothesis: unlike prior hybrid frameworks, which have only been validated on LLaMA-2, we evaluate a diverse

set of LLM backbones and observe consistently balanced validity across all of them, supporting the generalizability of the LLM-diffusion hybrid paradigm beyond any single backbone.
Finally, in terms of property statistics, recent denoising models such as ADiT, FlowMM, and MCFlow demonstrate strong performance, while approaches relying solely on LLMs (e.g., Crystal-Text-LLM) tend to underperform. Hybrid models (FlowLLM, CrysLLMGen) improve substantially over the LLM-only baseline, particularly on density. AdaCrysGen improves further still with its adaptive $\tau$ choice, narrowing the gap to the strongest denoising-based models. (Detailed analysis in D.1).

### 4.2 Generalization Across LLM Backbones

**Setup:** A central question for hybrid crystal generative models is whether their benefits are tied to a single LLM, or generalize across model families and scales. To test this, we repeat the de novo generation evaluation from Section 4.1 across five representative backbones spanning four LLM families: LLaMA-2 (7B), LLaMA-2 (13B), LLaMA-3 (8B), Gemma3 (12B), and Qwen3 (14B). For each backbone, we compare three configurations built on the identical LLM: the standalone LLM (Crystal-Text-LLM), the same LLM coupled with a fixed-timestep diffusion refinement stage (CrysLLMGen), and our adaptive-timestep variant (AdaCrysGen). We choose stability metrics for this comparison, and Table 2 reports Meta-Stable, MSUN, Stable, and SUN rates for all three configurations across all five backbones, alongside ADiT and MCFlow as representative standalone denoising baselines for reference.

| Models | Meta-Stable(%) | MSUN(%) | Stable(%) | SUN(%) |
|---|---|---|---|---|
| ADiT | 81.60 | 25.91 | 14.10 | 4.70 |
| MCFlow | 80.30 | 22.10 | 15.10 | 3.30 |
| Crystal-Text-LLM (LLaMA2-7B) | 56.60 | 26.66 | 12.67 | 4.84 |
| CrysLLMGen (LLaMA2-7B) | 62.02 | 35.94 | 16.79 | 9.21 |
| AdaCrysGen (LLaMA2-7B) | 71.80 | 45.68 | 18.64 | 10.68 |
| Crystal-Text-LLM (LLaMA2-13B) | 69.48 | 41.92 | 18.36 | 10.36 |
| CrysLLMGen (LLaMA2-13B) | 74.32 | 44.36 | 20.36 | 11.56 |
| AdaCrysGen (LLaMA2-13B) | 75.64 | <u>45.81</u> | 22.08 | 12.43 |
| Crystal-Text-LLM (LLaMA3-8B) | 78.92 | 40.96 | 21.76 | 10.44 |
| CrysLLMGen (LLaMA3-8B) | 82.00 | 43.12 | 22.28 | 11.72 |
| AdaCrysGen (LLaMA3-8B) | <u>83.78</u> | 44.84 | 24.52 | 12.64 |
| Crystal-Text-LLM (Gemma3-12B) | 78.76 | 40.56 | 24.20 | 11.44 |
| CrysLLMGen (Gemma3-12B) | 81.96 | 45.48 | <u>24.56</u> | 12.71 |
| AdaCrysGen (Gemma3-12B) | **87.92** | **48.27** | **26.36** | **14.16** |
| Crystal-Text-LLM (Qwen3-14B) | 71.88 | 37.72 | 16.64 | 7.68 |
| CrysLLMGen (Qwen3-14B) | 77.64 | 43.92 | 20.76 | 10.37 |
| AdaCrysGen (Qwen3-14B) | 78.94 | 45.12 | 22.16 | <u>13.88</u> |

Table 2: Stability metrics on the de novo generation task for standalone LLM (Crystal-Text-LLM), fixed-$\tau$ hybrid (CrysLLMGen), and adaptive-$\tau^*$ (AdaCrysGen) configurations across five LLM backbones, alongside ADiT and MCFlow. **Bold** denotes the best and <u>underline</u> denotes the second-best value.

**Results:** We observe a consistent trend: across all five backbones and all four stability metrics, performance increases monotonically from the standalone LLM to CrysLLMGen to AdaCrysGen, that is, AdaCrysGen $>$ CrysLLMGen $>$ Crystal-Text-LLM. This confirms that both the benefit of denoising-based refinement over standalone LLM generation, and the further benefit of replacing a fixed injection timestep with an adaptively predicted one, are not artifacts of a particular backbone; both hold uniformly across model families and scales. Notably, AdaCrysGen achieves its strongest absolute performance with Gemma3-12B rather than the largest backbone (Qwen3-14B), indicating that the framework's gains stem from the LLM-diffusion coupling itself rather than simply scaling up the underlying LLM. Taken together, these results support the generalizability of the LLM-diffusion hybrid paradigm, and of adaptive $\tau^*$ in particular, beyond any single backbone.

### 4.3 Text Conditioned Generation

**Setup:** One of the major advantages of using an LLM as the backbone of a material generation framework is its ability to incorporate user-defined prompts specifying desired conditions for the generated materials. However, existing approaches typically assume that conditioning information appears in a fixed order within the prompt; consequently, simply reordering these conditions at inference time can substantially degrade generation quality. To improve robustness to such prompt-order variations, we introduce two simple yet effective modifications to the LLM fine-tuning procedure: (1) an order-invariant training augmentation, and (2) an auxiliary condition-specific loss (Section 3.1).

Next, we evaluate the precision and robustness of AdaCrysGen on the conditional crystal generation task. We choose Crystal-Text-LLM (Gruver et al., 2024) and CrysLLMGen (Khastagir et al., 2026) as baselines for this comparison. For a fair comparison, we fix the LLM backbone of all three models to LLaMA-2 7B. Unlike prior works, which have only explored composition and space group number, we additionally evaluate the model's capacity on two further conditions: chemical formula and crystal system. Target conditions for a material are provided in the input prompt, which is then passed to the LLM component of each model, generating materials that aim to satisfy the given constraints.

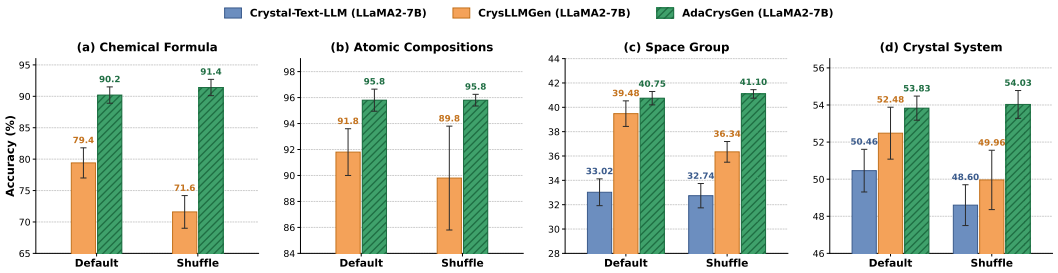



Figure 2: Results for text-conditioned generation with Crystal-Text-LLM, CrysLLMGen, and AdaCrysGen, all using the LLaMA-2 7B backbone, under the *Default* and *Shuffled* prompt-ordering setups. We report accuracy on four conditions – chemical formula, atomic composition, space group, and crystal system – with error bars denoting the corresponding variation across runs.

We evaluate the performance of all models under two prompt-ordering setups: (1) *Default*, where conditions appear in the fixed order formula, composition, space group, and crystal system; and (2) *Shuffled*, where this order is randomized independently for each sample. To assess the model's ability to generate crystal materials that satisfy specified conditions, we compare the generated outputs against the available ground-truth labels. For composition and formula, we parse the atomic formula from each generated CIF file and check for a match against the target composition and formula, respectively. Further, we use Pymatgen's SpacegroupAnalyzer (Ong et al., 2013), with a tolerance of 0.2 Å, to determine the space group and crystal system of both the generated and ground-truth materials, and compute the corresponding match rates.

**Results:** These experiments are conducted on 1,000 materials sampled from the MP-20 test set, and the results are presented in Figure 2. For atomic composition and chemical formula, we do not report Crystal-Text-LLM, since the composition predicted by Crystal-Text-LLM and CrysLLMGen is similar. We observe that AdaCrysGen outperforms CrysLLMGen across all four conditions in both the *Default* and *Shuffled* setups. This is because we train the LLM component of AdaCrysGen with an auxiliary condition-specific loss alongside its standard next-token prediction objective (Equation 1), which forces the LLM to better learn the conditions imposed in the prompt. Furthermore, in the *Shuffled* setup, the accuracy of CrysLLMGen degrades noticeably across all conditions relative to its own *Default*-setup accuracy, whereas AdaCrysGen remains largely stable. This is because, unlike CrysLLMGen, we shuffle the ordering of the conditioning sentences in each training prompt during LLM training, making AdaCrysGen less sensitive to prompt-ordering shuffles. Overall, these results show that our modifications to the LLM fine-tuning procedure not only enhance text-conditioned generation performance but also reduce sensitivity to prompt shuffling, addressing a key limitation of prior models.

## 5 Conclusion

We presented AdaCrysGen, which introduces two key novelties over prior hybrid LLM-diffusion frameworks: (i) a lightweight policy network that predicts a per-sample adaptive diffusion timestep $\tau_i^*$ in place of a single, globally fixed $\tau$; and (ii) an auxiliary condition-specific loss combined with order-invariant prompt augmentation to fine-tune the LLM component. With these novel modifications, AdaCrysGen achieves substantially more stable and valid crystals and is markedly more robust to conditional, text-guided generation. On the de novo generation task, AdaCrysGen consistently outperforms standalone diffusion, flow-matching, and LLM-based generators, as well as existing fixed-$\tau$ hybrid approaches, substantially increasing the fraction of generated materials that are stable, unique, and novel. These gains hold consistently across five LLM backbones spanning four model families, showing that adaptive-$\tau$ refinement is a general improvement rather than an artifact of a specific backbone. On text-conditioned generation, the same design further yields higher condition-satisfaction accuracy and markedly greater robustness to prompt-ordering shuffles than existing baselines. Together, these results establish the coupling point between a pretrained LLM and diffusion model as a meaningful, backbone-independent axis for improving hybrid crystal generation.

### AI USE STATEMENT

In this work, we did not use generative AI tools for research-related tasks, including ideation, experimental design, coding, or analysis. We used Large Language Models (Claude and ChatGPT) solely for editorial assistance, such as improving grammar, spelling, word choice, and the overall clarity of the manuscript. All AI-assisted revisions were carefully reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, and results presented in the paper.

### ETHICS STATEMENT

This work does not involve human subjects or personal data, and uses only the publicly available MP-20 dataset derived from the Materials Project. We are not aware of any elevated dual-use risk beyond that already present in the broader crystal-generation literature this work builds upon, nor of any conflicts of interest.

### REPRODUCIBILITY STATEMENT

We provide the full codebase and dataset as supplementary material, including detailed instructions for running the training and sampling pipelines and reproducing all reported results.

## REFERENCES

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.*, 31(9):3564–3572, 2019.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):1–8, 2021.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysxpp: An explainable property predictor for crystalline materials. *npj Computational Materials*, 8(1):43, 2022.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysmmnet: multimodal representation for crystal property prediction. In *Uncertainty in Artificial Intelligence*, pp. 507–517. PMLR, 2023a.

Kishalay Das, Bidisha Samanta, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysgnn: Distilling pre-trained knowledge to enhance property prediction for crystalline materials. *arXiv preprint arXiv:2301.05852*, 2023b.

Kishalay Das, Subhojyoti Khastagir, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Periodic materials generation using text-guided joint diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Daniel W Davies, Keith T Butler, Adam J Jackson, Jonathan M Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422, 2022.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Gautam R Desiraju. Cryptic crystallography. *Nature materials*, 1(2):77–79, 2002.

Mildred S Dresselhaus, Gene Dresselhaus, and Ado Jorio. *Group theory: application to the physics of condensed matter*. Springer Science & Business Media, 2007.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. The materials project: A materials genome approach to accelerating materials innovation, apl mater. *APL Mater.*, 2013.

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. *arXiv preprint arXiv:2309.04475*, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Reviews of Modern Physics*, 61(3):689–746, July 1989. ISSN 0034-6861. doi: 10.1103/revmodphys.61.689. URL http://dx.doi.org/10.1103/RevModPhys.61.689.

Chaitanya K Joshi, Xiang Fu, Yi-Lun Liao, Vahe Gharakhanyan, Benjamin Kurt Miller, Anuroop Sriram, and Zachary W Ulissi. All-atom diffusion transformers: Unified generative modelling of molecules and materials. In *International Conference on Learning Representations*, 2025.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysldm: Latent diffusion model for crystal material generation. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, et al. Llm meets diffusion: a hybrid framework for crystal material generation. *Advances in Neural Information Processing Systems*, 38:169190–169215, 2026.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, and Shuiwang Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21260–21287. PMLR, 23–29 Jul 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Steph-Yves Louis, Yong Zhao, Alireza Nasiri, Xiran Wang, Yuqi Song, Fei Liu, and Jianjun Hu. Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22(32):18141–18148, 2020.

Youzhi Luo, Chengkai Liu, and Shuiwang Ji. Towards symmetry-aware generation of periodic materials. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Jkc74vn1aZ.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching. *arXiv preprint arXiv:2406.04713*, 2024.

Shrimon Mukherjee, Kishalay Das, Partha Basuchowdhuri, Pawan Goyal, and Niloy Ganguly. Latent diffusion pretraining for crystal property prediction. *arXiv preprint arXiv:2606.00776*, 2026a.

Shrimon Mukherjee, Kishalay Das, Partha Basuchowdhuri, Pawan Goyal, and Niloy Ganguly. Model agnostic graph prompt learning for crystal property prediction. *arXiv preprint arXiv:2607.08996*, 2026b.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4(6), Jun 2020. ISSN 2475-9953. doi: 10.1103/physrevmaterials.4.063801. URL http://dx.doi.org/10.1103/PhysRevMaterials.4.063801.

Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nature Communications*, 16(1):4379, 2025.

Robert G. Parr and Yang Weitao. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, January 1995. ISBN 9780197560709. doi: 10.1093/oso/9780195092769.001.0001. URL http://dx.doi.org/10.1093/oso/9780195092769.001.0001.

Jonathan Schmidt, Love Pettersson, Claudio Verdozzi, Silvana Botti, and Miguel AL Marques. Crystal graph attention networks for the prediction of stable materials. *Science Advances*, 7(49): eabi7948, 2021.

Kiyoung Seong, Sungsoo Ahn, Sehui Han, and Changyoung Park. Multimodal crystal flow: Any-to-any modality generation for unified crystal modeling. *arXiv preprint arXiv:2602.20210*, 2026.

David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, March 2009. ISBN 9780470447710. doi: 10.1002/9780470447710. URL http://dx.doi.org/10.1002/9780470447710.

Anuroop Sriram, Benjamin Miller, Ricky TQ Chen, and Brandon Wood. Flowllm: Flow matching for material generation with large language models as base distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046, 2024.

H Wu, Y Song, J Gong, Z Cao, Y Ouyang, J Zhang, H Zhou, WY Ma, and J Liu. A periodic bayesian flow for material generation (2025). *arXiv preprint arXiv:2502.02016*, 2025.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120(14):145301, 2018.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *The 36th Annual Conference on Neural Information Processing Systems*, 2022.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BnQY9XiRAS.

Anthony Zee. *Group theory in a nutshell for physicists*, volume 17. Princeton University Press, 2016.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

# Technical Appendix of AdaCrysGen

## A Limitations and Scope for Future Work

Our approach has few limitations, which we leave for future work.

First, our reward signal treats stability, as estimated via CHGNet-relaxed $E_{\text{hull}}$, as a proxy for true synthesizability. While this proxy is computationally tractable at scale, CHGNet is itself a machine-learned surrogate for DFT, and stability alone does not guarantee that a generated structure is experimentally realizable. This work is a stepping stone toward exploring adaptive LLM-diffusion fusion, and we obtain promising results with this reward signal; however, designing a stronger reward signal and incorporating more accurate approximations to DFT are promising directions for future work.

Second, the policy network $\pi_\phi$ selects $\tau^*$ from a discrete, precomputed candidate set $\mathcal{T}$ rather than predicting a continuous timestep, which caps the granularity of the diffusion injection point that can be selected for any given sample. Extending the policy to predict a continuous timestep is a direction for future work.

Lastly, among denoising generative models, we adopt diffusion-based models, as they have shown more promising results than their flow-based counterparts for material generation. As better generative models emerge, integrating LLMs with them in a similarly adaptive fashion is an interesting direction for future work.

## B Related Work

Crystal structure generation has been substantially advanced by diffusion-based generative models, which achieve higher structural validity and thermodynamic stability than earlier generative approaches. A central driver of this progress is the incorporation of equivariance: since periodic crystals are invariant under rotations, translations, and permutations of symmetric atoms, models built on equivariant graph neural networks are particularly well suited to this domain. Early instances of this paradigm, including CDVAE (Xie et al., 2021) and SyMat (Luo et al., 2023), combine a variational autoencoder with a score-based denoising network operating on atomic coordinates, relying on equivariant message passing to respect Euclidean and periodic symmetries. Later work, notably DiffCSP (Jiao et al., 2023) and MatterGen (Zeni et al., 2023), moves to a joint diffusion process that generates atomic composition, fractional coordinates, and lattice parameters together within a single framework, whereas DiffCSP++ (Jiao et al., 2024) and SymmCD (Levy et al., 2025) instead exploit space-group symmetry to guide material generation. Because operating directly in this high-dimensional, multimodal feature space incurs substantial training and inference cost, recent works such as CrysLDM (Khastagir et al., 2025) and ADiT (Joshi et al., 2025) instead perform diffusion in a compact latent space, substantially reducing sampling time and resource consumption relative to feature-domain diffusion models.

As an alternative to diffusion, flow matching (FM) (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023) has recently gained traction for material generation. Rather than iteratively denoising samples from a Gaussian prior, FM learns a time-dependent velocity field that continuously transports an arbitrary base distribution toward the target distribution of stable crystals. FlowMM (Miller et al., 2024) was the first to instantiate this framework for materials, using a representation that respects global rotational and translational symmetry together with periodic boundary conditions, and equivariant flows to construct symmetry-invariant base distributions. In a related direction, CrysBFN (Wu et al., 2025) adopts periodic Bayesian Flow Networks with entropy conditioning and non-monotonic dynamics to better capture periodicity in non-Euclidean space, improving both sampling efficiency and generation quality.

Lately, text-guided diffusion models such as TGDMat (Das et al., 2025) and Chameleon (Park et al., 2025) integrate contextual embeddings from pretrained models like MatSciBERT or CLIP into GNN-based denoising networks to generate stable periodic materials aligned with textual descriptions. Large language models (LLMs) have also gained substantial traction for material generation, owing to their ability to process natural language prompts and enable conditional generation. By representing crystal structures as textual descriptions of unit cells and atomic positions, token-based language models (Antunes et al., 2024; Gruver et al., 2024) achieve strong performance in generating stable

and valid materials. However, these models are effective at modeling discrete atom types but struggle to model continuous atomic coordinates and lattice parameters. Hybrid approaches address this limitation by pairing LLMs with generative refinement: FlowLLM (Sriram et al., 2024), as discussed above, derives its base distribution from an LLM, while CrysLLMGen (Khastagir et al., 2026) couples an LLM with a diffusion model, using the LLM to predict atom types and composition and the diffusion component to refine the continuous fractional coordinates and lattice parameters.

Another line of work in material design targets fast and accurate crystal property prediction. Towards this goal, graph neural network (GNN)-based approaches (Xie & Grossman, 2018; Chen et al., 2019; Choudhary & DeCost, 2021; Das et al., 2023a; Louis et al., 2020; Park & Wolverton, 2020; Schmidt et al., 2021; Mukherjee et al., 2026b) have emerged as powerful models for learning robust representations of crystal materials, improving property prediction precision. More recently, graph transformer-based architectures (Yan et al., 2022; Lin et al., 2023; Yan et al., 2024) have been proposed to learn periodic graph representations of materials, improving performance while also being substantially faster than prior state-of-the-art models. Moreover, the scarcity of labeled data makes these models difficult to train across all properties; to mitigate this, recent studies (Das et al., 2022; 2023b; Mukherjee et al., 2026a) have shown promising results using transfer learning, pre-training, and knowledge distillation.

# C Preliminaries

## C.1 Crystal Structure Representation and Symmetries

A crystal can be represented as a periodic three-dimensional arrangement of atoms, where a finite unit cell is repeatedly translated along three lattice directions to construct the complete structure. For a unit cell containing $N$ atoms, we represent the crystal using three components: the atom types, atomic coordinates, and lattice vectors. The atom identities are encoded by the atom type matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]^T \in \mathbb{R}^{N\times k}$, where each $a_i$ is a one-hot representation of the atom type and $k$ denotes the number of possible atomic species. The atomic positions are represented by the Coordinate Matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N\times 3}$, where $x_i \in \mathbb{R}^3$ gives the position of the $i^{th}$ atom within the unit cell. The periodicity of the crystal is defined by the Lattice Matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]^T \in \mathbb{R}^{3\times 3}$, whose vectors determine the three directions along which the unit cell is repeated. Thus, a crystal can be represented as $\mathbf{C} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$. Its corresponding infinite periodic structure is obtained by translating each atom over integer combinations of the lattice vectors, while preserving the atomic identities:

$$\mathbf{X}' = \{\mathbf{x}'_i \mid \mathbf{x}'_i \mid= \mathbf{x}_i + \sum_{j=1}^{3} k_j \mathbf{l}_j,\ k_j \in \mathbb{Z}\}, \qquad \mathbf{A}' = \{\mathbf{a}'_i \mid \mathbf{a}'_i = \mathbf{a}_i\}, \qquad i = 1, \ldots, N.$$

## C.2 Symmetry in Crystal Structure

Generative modeling of crystalline materials aims to learn the probability distribution $p(\mathbf{M})$ over crystal structures. Since crystals obey several fundamental physical symmetries (Dresselhaus et al., 2007; Zee, 2016), the learned distribution should remain unchanged under equivalent representations of the same material. In particular, crystal generation should respect invariance to permutation, translation, rotation, and periodic transformations.

- **Permutation Invariance:** The physical identity of a crystal does not depend on the ordering used to index its atoms. Consider a crystal $\mathbf{M} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$ and a permutation matrix $\mathbf{P}$. Reordering the atom types and coordinates gives
$$\mathbf{M}_P = (\mathbf{PA}, \mathbf{PX}, \mathbf{L}).$$
Since $\mathbf{M}$ and $\mathbf{M}_P$ represent the same crystal, the learned distribution should satisfy
$$p(\mathbf{M}) = p(\mathbf{M}_P).$$

- **Translation Invariance:** A global translation of all atoms does not modify the underlying crystal structure. For any translation vector $\mathbf{u} \in \mathbb{R}^3$, let
$$\mathbf{M}_T = (\mathbf{A}, \mathbf{X} + \mathbf{1}\mathbf{u}^T, \mathbf{L}).$$
The translated structure is physically equivalent to the original one, and therefore
$$p(\mathbf{M}) = p(\mathbf{M}_T).$$

- **Rotational Invariance:** The representation of a crystal should also be invariant to a global rotation of the structure. Given an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{3\times3}$ satisfying $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, rotating both the atomic positions and lattice vectors produces an equivalent representation,

$$\mathbf{M}_R = (\mathbf{A}, \mathbf{X}\mathbf{Q}^T, \mathbf{L}\mathbf{Q}^T).$$

  Hence, the corresponding probabilities should remain identical:

$$p(\mathbf{M}) = p(\mathbf{M}_R).$$

- **Periodic Invariance:** Because a crystal is formed by periodically repeating its unit cell along the lattice directions, the same material can admit multiple equivalent coordinate representations. Let $\mathbf{K} \in \mathbb{Z}^{N\times3}$ denote an integer-valued matrix specifying lattice translations for individual atoms. The transformed coordinates

$$\mathbf{X}' = \mathbf{X} + \mathbf{K}\mathbf{L}$$

  describe a periodically equivalent structure. Thus,

$$\mathbf{M}' = (\mathbf{A}, \mathbf{X}', \mathbf{L})$$

  represents the same crystal as $\mathbf{M}$, and the model should satisfy

$$p(\mathbf{M}) = p(\mathbf{M}').$$

# D Additional Experiments

## D.1 Benchmarking on De Novo Material Generation (DNG) Task

We report the full De Novo Generation results in Table 1 of the main manuscript, which includes all baselines alongside results for 5 representative backbones of AdaCrysGen. In Table 3, we report the complete set of all 14 LLM backbones spanning different model families and scales for AdaCrysGen. Overall, every variant of AdaCrysGen surpasses existing baselines on most metrics, achieving superior performance across the majority of models compared in Table 3. We highlight several key takeaways from the results below.

**Stability Metrics.** In terms of stability, all variants of AdaCrysGen excel at generating unique and novel materials, substantially outperforming standalone denoising-based generative models, including both diffusion and flow-matching approaches, on the stricter MSUN, Stable, and SUN metrics. On the less strict Meta-Stable metric, most AdaCrysGen backbones remain competitive with or exceed the strongest denoising baselines, with the LLaMA-3 and Gemma3-12B variants surpassing even ADiT, the strongest standalone denoising model.

We further compare AdaCrysGen against standalone denoising-based crystal generators, including diffusion models such as CDVAE, DiffCSP, DiffCSP++, MatterGen, SymmCD, and ADiT, as well as flow-matching approaches such as FlowMM and MCFlow. Among the diffusion-based approaches, ADiT is the strongest baseline on Meta-Stable and Stable, while DiffCSP achieves the highest MSUN and MatterGen achieves the highest SUN. Compared with the strongest diffusion baseline on each respective metric, the Gemma3-12B variant of AdaCrysGen improves Meta-Stable, MSUN, Stable, and SUN by $7.7\%$, $50.0\%$, $87.0\%$, and $99.4\%$, respectively. These improvements are particularly substantial on the stricter Stable and SUN criteria, with AdaCrysGen achieving nearly twice the SUN rate of the strongest diffusion-based baseline. The remaining AdaCrysGen variants are also competitive with all standalone denoising-based crystal generators, outperforming most of them. On the stricter MSUN, Stable, and SUN metrics, every AdaCrysGen backbone surpasses every diffusion- and flow-matching-based baseline, while on Meta-Stable, all variants remain ahead of the majority of denoising baselines.

A similar trend is observed against flow-matching approaches. Compared with MCFlow, the stronger of the two flow-matching baselines overall, the Gemma3-12B variant of AdaCrysGen improves Meta-Stable by $9.5\%$, MSUN by $118.4\%$, Stable by $74.6\%$, and SUN by $329.1\%$. The particularly large improvement in SUN indicates that the higher Meta-Stability achieved by AdaCrysGen translates much more effectively into generated materials that are simultaneously stable, unique, and novel, rather than producing a large number of metastable structures that fail the stricter uniqueness and novelty criteria.

| Models | Validity ↑ | | Property ↓ | | Stability, Uniqueness and Novelty ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| | Comp (%) | Struc (%) | Density | Element | M-Stable (%) | MSUN (%) | Stable (%) | SUN (%) |
| **Diffusion Models** | | | | | | | | |
| CDVAE | 86.70 | **100** | 0.687 | 1.432 | 23.58 | 21.99 | 3.08 | 2.56 |
| DiffCSP | 83.25 | **100** | 0.352 | 0.339 | 35.04 | 32.19 | 7.36 | 5.61 |
| DiffCSP++ | 85.12 | 99.94 | 0.235 | 0.375 | 42.39 | 30.56 | 8.58 | 6.55 |
| MatterGen | 86.34 | **100** | 0.459 | 0.254 | 78.00 | 21.00 | 13.00 | 7.10 |
| SymmCD | 87.13 | 92.30 | 0.531 | 0.210 | 40.01 | 31.69 | 9.99 | 6.76 |
| ADiT | 90.76 | 99.34 | **0.076** | 0.056 | 81.60 | 25.91 | 14.10 | 4.70 |
| **Flow Matching Models** | | | | | | | | |
| FlowMM | 83.19 | 96.85 | 0.239 | 0.083 | 31.64 | 22.46 | 4.76 | 3.06 |
| MCFlow | 89.85 | 99.44 | 0.136 | **0.055** | 80.30 | 22.10 | 15.10 | 3.30 |
| **LLM-Only Models** | | | | | | | | |
| Crystal-Text-LLM (LLaMA2-7B) | 91.19 | 97.70 | 1.575 | 0.271 | 56.60 | 26.66 | 12.67 | 4.84 |
| **Hybrid Models (Fixed $\tau$)** | | | | | | | | |
| FlowLLM (LLaMA2-7B) | 91.19 | 99.94 | 1.142 | 0.271 | 60.19 | 30.17 | 14.66 | 6.97 |
| CrysLLMGen (LLaMA2-7B) | 91.19 | **100** | 1.105 | 0.271 | 62.02 | 35.94 | 16.79 | 9.21 |
| **Ours: AdaCrysGen** | | | | | | | | |
| AdaCrysGen (LLaMA2-7B) | 91.19 | **100** | 1.0655 | 0.271 | 71.80 | 45.68 | 18.64 | 10.68 |
| AdaCrysGen (LLaMA2-7B-chat) | 90.95 | **100** | 1.0975 | 0.243 | 74.92 | 45.88 | 21.84 | 12.36 |
| AdaCrysGen (LLaMA2-13B) | 91.96 | **100** | 1.0640 | 0.238 | 75.64 | 45.81 | 22.08 | 12.43 |
| AdaCrysGen (LLaMA2-13B-chat) | 91.08 | **100** | 1.3731 | 0.2213 | 79.82 | 46.61 | 23.17 | 13.67 |
| AdaCrysGen (LLaMA3-8B) | 93.44 | **100** | 1.0205 | 0.185 | 83.78 | 44.84 | 24.52 | 12.64 |
| AdaCrysGen (LLaMA3-8B-Instruct) | 92.68 | **100** | 1.2943 | 0.1802 | 83.16 | 45.64 | 23.04 | 11.52 |
| AdaCrysGen (Gemma3-1B) | 90.41 | **100** | 1.6385 | 0.1748 | 77.08 | 46.68 | 22.68 | 12.92 |
| AdaCrysGen (Gemma3-4B) | 91.52 | **100** | 1.2512 | 0.1773 | 79.52 | 47.21 | 23.38 | 13.75 |
| AdaCrysGen (Gemma3-12B) | 91.64 | **100** | 1.0672 | 0.192 | **87.92** | **48.27** | **26.36** | **14.16** |
| AdaCrysGen (Qwen3-0.6B) | 92.02 | **100** | 0.8278 | 0.2516 | 75.72 | 43.68 | 19.44 | 9.98 |
| AdaCrysGen (Qwen3-1.7B) | **94.13** | **100** | 0.7576 | 0.233 | 75.12 | 45.36 | 16.28 | 8.56 |
| AdaCrysGen (Qwen3-4B) | 92.76 | **100** | 0.9548 | 0.2067 | 78.20 | 43.04 | 19.56 | 11.64 |
| AdaCrysGen (Qwen3-8B) | 90.20 | **100** | 0.8831 | 0.2378 | 78.68 | 44.32 | 21.68 | 12.52 |
| AdaCrysGen (Qwen3-14B) | 91.12 | **100** | 0.5647 | 0.231 | 78.94 | 45.12 | 22.16 | 13.88 |

Table 3: Benchmarking results on the *De Novo Generation Task* across different classes of generative models, including diffusion models, flow-matching models, LLM-only models, fixed-$\tau$ hybrid models, and our adaptive hybrid framework. Results are reported across validity, property statistics, and stability/uniqueness/novelty metrics. **Bold** denotes the best value and underline denotes the second-best value in each column.

Next, we compare AdaCrysGen against Crystal-Text-LLM, which directly generates crystal structures using a LLaMA-2 7B LLM. Using the same LLaMA-2 7B backbone, AdaCrysGen improves Meta-Stable, MSUN, Stable, and SUN by $26.9\%$, $71.3\%$, $47.1\%$, and $120.7\%$, respectively, with the largest gain on SUN, where AdaCrysGen more than doubles the pure LLM baseline. This highlights the benefit of refining the initial LLM-generated structures, substantially increasing the fraction of generated materials that simultaneously satisfy stability, uniqueness, and novelty.

Finally, we compare AdaCrysGen against existing hybrid approaches, including FlowLLM and CrysLLMGen, which combine an LLM with a subsequent generative refinement stage. Using the same LLaMA-2 7B backbone, AdaCrysGen improves over FlowLLM by $19.3\%$, $51.4\%$, $27.1\%$, and $53.2\%$ on Meta-Stable, MSUN, Stable, and SUN, respectively. More importantly, compared with CrysLLMGen, which also uses the same LLaMA-2 7B backbone but applies a single, globally selected diffusion timestep to all generated samples, AdaCrysGen improves these same metrics by $15.8\%$, $27.1\%$, $11.0\%$, and $16.0\%$, respectively. Since the LLM backbone is held fixed across these comparisons, the consistent gains demonstrate the benefit of sample-adaptive denoising over a fixed, sample-agnostic refinement strategy applied uniformly across generated crystals.

**Validity Metrics.** On validity metrics, similar to CrysLLMGen, all variants of AdaCrysGen achieve balanced structural and compositional validity. Denoising-based models (diffusion or flow-matching-based models) are generally effective at modeling atomic coordinates and lattice parameters, and

**Algorithm 1** Adaptive Crystal Sampling with AdaCrysGen

**Require:** Pretrained LLM; pretrained diffusion model $f_\theta^{\text{Diff}}$; trained timestep policy $\pi_\phi$; candidate timesteps $\mathcal{T} = \{\tau_1, \ldots, \tau_B\}$; text prompt $\mathcal{P}$
**Ensure:** Generated crystal $\mathbf{C}_{\text{new}}$

**Stage I: LLM-based Crystal Initialization**
$(\hat{\mathbf{A}}, \hat{\mathbf{X}}, \hat{\mathbf{L}}) \sim \text{LLM}(\mathcal{P})$
$\mathbf{A}_{\text{new}} \leftarrow \hat{\mathbf{A}}$ ▷ atomic composition remains fixed

**Stage II: Adaptive Timestep Selection**
$\mathbf{c} \leftarrow \text{FEATURIZE}(\hat{\mathbf{A}}, \hat{\mathbf{X}}, \hat{\mathbf{L}})$ ▷ Sec. 3.3
$\tau^* \leftarrow \arg\max_{\tau \in \mathcal{T}} \pi_\phi(\tau \mid \mathbf{c})$

**Stage III: Diffusion-based Structural Refinement**
$(\mathbf{X}_{\text{new}}, \mathbf{L}_{\text{new}}) \leftarrow \text{DIFFUSION-REFINE}\big(\hat{\mathbf{A}}, \hat{\mathbf{X}}, \hat{\mathbf{L}}, \tau^*; f_\theta^{\text{Diff}}\big)$ ▷ reverse diffusion from $\tau^*$ to 0

$\mathbf{C}_{\text{new}} \leftarrow (\mathbf{A}_{\text{new}}, \mathbf{X}_{\text{new}}, \mathbf{L}_{\text{new}})$
**return** $\mathbf{C}_{\text{new}}$

consequently achieve high structural validity, but tend to struggle relatively more with atomic composition, resulting in generally lower compositional validity. LLMs exhibit the opposite pattern: strong compositional validity ($\geq 90\%$) but weaker structural validity, reflecting the difficulty of generating precise continuous geometry through autoregressive text generation. Hybrid models such as FlowLLM and CrysLLMGen address this trade-off by combining an LLM for composition with a denoising-based refinement stage for structure, achieving strong performance on both fronts simultaneously. AdaCrysGen reinforces this hypothesis: unlike prior hybrid frameworks, which have only been validated on LLaMA-2, we evaluate a diverse set of LLM backbones and observe consistently balanced validity across all of them, supporting the generalizability of the LLM-diffusion hybrid paradigm beyond any single backbone.

**Property Statistics.** In terms of property statistics, recent denoising-based models such as ADiT, FlowMM, and MCFlow demonstrate strong performance, whereas approaches relying solely on LLMs (e.g., Crystal-Text-LLM) generally underperform. Hybrid approaches such as FlowLLM and CrysLLMGen substantially improve over the LLM-only baseline, particularly in terms of density ($d_\rho$). AdaCrysGen further improves density matching through its adaptive choice of $\tau$, narrowing the gap to the strongest denoising-based models across most LLM backbones. For $d_{\text{elem}}$, Crystal-Text-LLM, FlowLLM, CrysLLMGen, and AdaCrysGen achieve equivalent results because all these methods retain the atom types generated by the LLM backbone; consequently, the elemental composition is unchanged by the subsequent refinement stage.