SYNERGISTIC MULTI-TASK LEARNING FOR ELECTRONIC DENSITY OF STATES PREDICTION

Kunmin Jang[1*] **Dongik Park**[1*] **Jaewon Bae**[1] **Chanyoung Park**[2†]
[1]NanoForge AI, Seoul, Republic of Korea
[2]Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{alexjang, dongik.park, jaewon.bae}@nanoforgeai.com
cy.park@kaist.ac.kr

ABSTRACT

First-principles calculations provide detailed electronic structure information but are computationally expensive, limiting their application to large-scale materials screening. Machine learning offers a promising alternative, yet existing approaches typically predict density of states (DOS), element-projected DOS (EPDOS), Fermi level, and band gap using separate models, missing potential synergies between these interrelated quantities. We propose DOSForge, a multi-task architecture that exploits such synergies via cross-conditioning between decoders and parameter-free band gap supervision. Experiments show consistent gains across all four targets (DOS, EPDOS, Fermi level, and band gap), whereas naive multi-task baselines exhibit clear trade-offs across targets. DOSForge achieves state-of-the-art DOS prediction and demonstrates that principled multi-task design can turn competing objectives into mutual gains.

1 INTRODUCTION

Understanding the electronic structure of crystalline materials is essential for predicting their functional properties. First-principles methods such as density functional theory (DFT) (Kohn & Sham, 1965) provide rigorous electronic structure calculations, but their computational cost limits their application to large-scale materials discovery. Machine learning has emerged as a promising alternative, enabling rapid prediction of material properties directly from crystal structures, from scalar quantities like band gap (Dunn et al., 2020) to full spectral outputs like density of states (Bai et al., 2022; Lee et al., 2023; Xie et al., 2025).

Among various electronic-structure descriptors, we focus on the following interrelated quantities. The density of states (DOS) describes the distribution of electronic states across energy levels, providing direct insight into band gaps, metallicity, and transport properties. Element-projected DOS (EPDOS) decomposes this spectrum by atomic species, revealing which elements contribute at each energy level. This decomposition is critical for understanding bonding character and guiding compositional design. The Fermi level sets the reference for electronic state occupation, while the band gap largely governs optical and electronic behavior. These quantities arise from the same underlying electronic structure and satisfy known physical relationships: element-wise contributions ideally sum to the total DOS, and the band gap corresponds to the zero-density region around the Fermi level.

Existing machine learning approaches typically predict these quantities using separate models or derive secondary properties from predicted spectra via post-processing steps. While some methods acknowledge relationships between the quantities, prior architectures have not been designed to exploit multi-task synergies, and simply adding auxiliary objectives can hurt performance (Xie et al., 2025).This often results in trade-offs where improving one task degrades another. We propose DOSForge (Figure 1), a multi-task architecture that addresses this challenge through cross-conditioning and parameter-free band gap supervision.

[*]Both authors contributed equally to this research.
[†]Corresponding author.

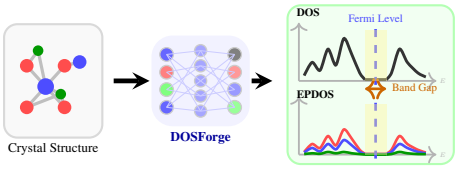

Figure 1: DOSForge predicts DOS, EPDOS, Fermi level, and band gap from crystal structure.

Rather than decoding DOS and EPDOS independently, each decoder receives intermediate representations from the other during decoding. This design is motivated by complementary information flow: element-level features help compose the DOS by aggregating atomic contributions, while global spectral context constrains EPDOS by enforcing consistency with the overall electronic structure. The band gap is extracted directly from the predicted DOS, avoiding additional learnable parameters that could conflict with spectral prediction. The Fermi level is predicted via attention pooling over encoder outputs. Experiments demonstrate that DOSForge achieves state-of-the-art DOS prediction while jointly predicting EPDOS, Fermi level, and band gap in a physically consistent manner.

2 Related Work

Graph neural networks for materials have advanced from predicting scalar properties (Xie & Grossman, 2018; Choudhary & DeCost, 2021) to full spectral quantities such as DOS. Mat2Spec (Kong et al., 2022) encodes crystal graphs into probabilistic embeddings and decodes DOS via contrastive learning. Xtal2DoS (Bai et al., 2022) improves upon this by introducing cross-attention, where learnable energy queries attend to atomic embeddings. DOSTransformer (Lee et al., 2023) extends this formulation by predicting DOS and then estimating band gap and Fermi level from the predicted DOS, but these secondary predictions rely on post-processing models rather than being trained end-to-end. EqDOS (Xie et al., 2025) predicts atom and orbital-projected DOS using equivariant networks, but does not perform fully joint optimization of these targets; it also reports that adding auxiliary objectives can degrade DOS accuracy, highlighting the difficulty of effective multi-task learning. DOSForge addresses this challenge through structured decoder interactions, enabling joint prediction with mutual benefit.

3 Methodology

3.1 Problem Formulation

Given a crystal structure $M = (A, P, L)$ (Yan et al., 2024) with atomic features $A \in \mathbb{R}^{d_a \times n}$ for n atoms, positions $P \in \mathbb{R}^{3\times n}$, and lattice matrix $L \in \mathbb{R}^{3\times 3}$, we jointly predict:

- DOS $\mathbf{d} \in \mathbb{R}^{T}_{\geq 0}$ on a uniform grid of T=256 bins over $[-10, 10]$ eV relative to the Fermi level
- Element-projected DOS $\{\mathbf{d}_\ell \in \mathbb{R}^{T}_{\geq 0}\}_{\ell\in\mathcal{E}}$, also on a uniform grid of T=256 bins, where $\mathcal{E}$ is the set of element types present in the structure
- Fermi level $E_F \in \mathbb{R}$ and band gap $E_g \in \mathbb{R}_{\geq 0}$

These quantities are physically related: *EPDOS* contributions ideally sum to *DOS*, both represented relative to the *Fermi level*, and *band gap* corresponds to the zero-density region in *DOS* around the *Fermi level*. Our goal is to design an architecture that exploits these relationships to enable effective joint prediction.

3.2 Cross-Conditioning

Standard multi-task learning uses shared encoders with independent prediction heads, enabling only implicit information sharing. To enable direct information exchange between DOS and EPDOS predictions, we explore different coupling strategies between their decoders (Figure 2). In (a), both outputs are predicted in parallel with no coupling. In (b), EPDOS is aggregated into DOS, providing bottom-up information flow; this improves DOS prediction. In (c), DOS informs element decomposition via top-down flow; this improves EPDOS prediction. Observing that (b) and

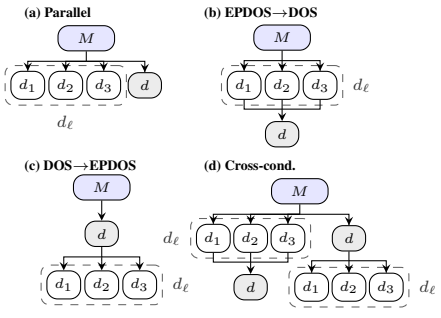

Figure 2: Information flow variants.

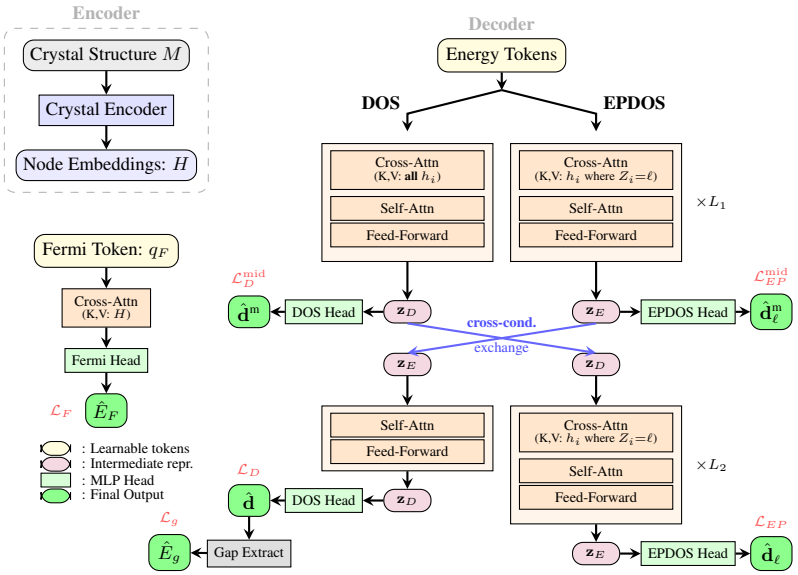

Figure 3: DOSForge architecture. After L_1 blocks, DOS and EPDOS streams exchange representations before branch-specific L_2 blocks. Intermediate outputs ($\hat{\mathbf{d}}^{\text{m}}$, $\hat{\mathbf{d}}_\ell^{\text{m}}$) are supervised at the exchange point.

(c) each benefit different tasks, we design cross-conditioning (d) to achieve gains on both. In this unified architecture, both decoders mutually exchange intermediate representations, integrating bottom-up element aggregation with top-down spectral guidance.

3.3 DOSForge Architecture

Figure 3 shows the DOSForge architecture. Crystal structures are encoded using iComFormer (Yan et al., 2024), a graph neural network that constructs E(3) invariant complete graph representations, producing atomic embeddings $H = \{h_i\}_{i=1}^n$ where $h_i \in \mathbb{R}^d$. Spectra are decoded using $N_p{=}32$ learnable energy patch tokens that attend to these embeddings via sigmoid-gated cross-attention. For DOS, all atoms serve as keys and values; for EPDOS of element ℓ, only atoms of that element are included. After L_1 transformer blocks, the two streams exchange intermediate representations: element spectra aggregate into a DOS initializer via scatter-sum along elements, while the DOS representation broadcasts to each element group. Each stream then continues decoding for L_2 blocks. The Fermi level is predicted via attention pooling over encoder outputs. Architecture and training details are provided in Appendix A and B.

3.4 Band Gap Extraction with Parameter-free Supervision

Predicting the band gap with a separate learnable head can cause gradient conflicts with DOS prediction, as both objectives compete for the shared encoder. We instead extract the band gap directly from the predicted DOS, introducing no additional learnable parameters, so that band gap supervision regularizes the DOS rather than competing with it. While the band gap is well-defined physically, its conventional extraction from a discretized DOS relies on non-differentiable operations such as indicator functions and hard edge detection. We address this via **soft edge detection through cumulative integration**. Starting from the Fermi level, we accumulate the DOS outward into both left (L; decreasing energy) and right (R; increasing energy) directions. Within the gap, the running integral stays near zero; once band states are reached, it increases rapidly.

Let $\mathbf{d}_L$ and $\mathbf{d}_R$ denote DOS values on energy bins ordered outward from E_F to lower and higher energies, respectively. We approximate the gap as

$$\hat{E}_g = \left[\sum_i \text{ReLU}(1 - k \cdot \text{cumsum}(\mathbf{d}_L)_i) + \sum_j \text{ReLU}(1 - k \cdot \text{cumsum}(\mathbf{d}_R)_j)\right] \Delta E, \quad (1)$$

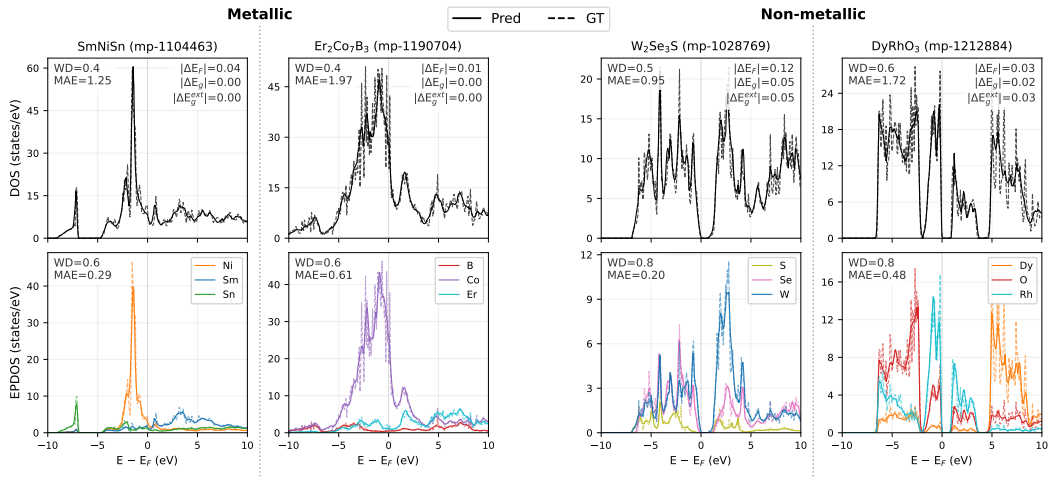

Figure 4: Selected test-set predictions for metallic (left) and non-metallic (right) materials. Solid: prediction, dashed: DFT. Panels report WD, MAE, and errors for E_F, E_g, E_g^{ext}.

where $\text{cumsum}(\mathbf{d})_i := \sum_{t=1}^{i}(\mathbf{d})_t$ is a cumulative sum from E_F outward, ΔE is the bin width, and k controls the transition sharpness. This extraction enables gradients from the band gap loss to flow through the predicted DOS to the shared encoder (details in Appendix A.4).

4 Experiments

4.1 Experimental Setup

Dataset. We use 62,964 structures from the Materials Project (Jain et al., 2013) spanning 84 elements, excluding noble gases. Data is split 7 : 1.5 : 1.5 for train/validation/test, stratified by point group. DOS and EPDOS are aligned to the Fermi level and interpolated onto a uniform grid of T=256 bins over $[-10, 10]$ eV using PCHIP (Fritsch & Carlson, 1980) to preserve spectral shape.

Baselines. Baselines include Xtal2DoS (Bai et al., 2022), DOSTransformer (Lee et al., 2023), and EqDOS (Xie et al., 2025), all trained using publicly available code. For ablations, variants include single-task models, parallel multi-task (shared encoder, separate heads), and the four conditioning configurations from Figure 2.

Metrics. For spectral outputs (DOS, EPDOS), we use Wasserstein distance (WD) as the primary metric. DOS and EPDOS exhibit many sharp peaks, making point-wise metrics like MAE overly sensitive to small peak position shifts even when the overall spectral shape is preserved. WD measures the minimum cost to transform one distribution into another, making it more robust to small peak misalignments while capturing global shape similarity. Scalar targets (Fermi level, band gap) are evaluated using MAE. We additionally report E_g^{ext} (MAE), computed by a simple non-differentiable threshold-based edge scan (Appendix A.4) to measure how well the predicted DOS exhibits a clear gap.

4.2 DOS Prediction

Table 1 compares DOSForge against prior DOS prediction methods. DOSForge achieves state-of-the-art performance, reducing Wasserstein distance by 11.6% relative to the best baseline (DOSTransformer: 6.45 → 5.70) and MAE by 13.4% (3.20 → 2.77). This improvement comes from multi-task learning with cross-conditioning, which we analyze in the next section. Figure 4 shows selected predictions; Appendix C provides examples across performance percentiles.

Table 1: DOS prediction performance.

Method	MAE	MSE	R^2	WD
Xtal2DoS	3.24	83.2	.475	6.82
DOSTransformer	3.20	82.2	.463	6.45
EqDOS	3.26	97.3	.342	10.4
DOSForge	**2.77**	**82.1**	**.495**	**5.70**

Table 2: Effect of task scope and conditioning direction. Single: separate per-task models. Parallel: shared encoder, independent heads. Best in **bold**. E_g^{ext}: band gap extracted from predicted DOS.

Tasks	Model	DOS (WD ↓)	EPDOS (WD ↓)	E_F (MAE ↓)	E_g (MAE ↓)	E_g^{ext} (MAE ↓)
1-task	Single	6.543	8.401	0.229	0.189	0.284[†]
2-task	Parallel	6.080	8.746	–	–	0.261
	EPDOS→DOS	5.788	8.709	–	–	0.247
	DOS→EPDOS	6.290	8.293	–	–	0.257
	Cross-cond.	5.783	7.932	–	–	0.232
4-task	Parallel	6.218	8.889	**0.211**	0.178	0.240
	DOSForge[‡]	5.733	7.953	0.217	0.192	0.196
	DOSForge	**5.702**	**7.835**	0.217	**0.171**	**0.171**

[†]Extracted from DOS (Single). [‡]Learned MLP head for band gap instead of extraction from predicted DOS.

4.3 Multi-task Synergy Analysis

Table 2 compares single-task baselines, 2-task variants (DOS + EPDOS), and 4-task joint prediction. Single combines results from four separate per-task models. Parallel uses a shared encoder with independent heads; the band gap is predicted via learned attention pooling from the encoder, using the same head design as for the Fermi level. DOSForge uses cross-conditioning and parameter-free band gap extraction; DOSForge[‡] replaces the latter with a learned MLP head.

Cross-conditioning enables mutual improvement. We first analyze the conditioning direction (Figure 2a~d) in the 2-task setting. (a) Parallel improves DOS over Single but degrades EPDOS, revealing a trade-off in naive multi-task learning. Unidirectional conditioning shows that (b) EPDOS→DOS significantly improves DOS (6.54 → 5.79) but worsens EPDOS, while (c) DOS→EPDOS improves EPDOS (8.40 → 8.29) but DOS falls below Parallel. Each direction benefits primarily the receiving task, leaving the other unchanged or degraded. (d) Cross-conditioning resolves this trade-off, achieving the best among all 2-task methods and substantially improving both DOS (6.54 → 5.78) and EPDOS (8.40 → 7.93).

4-task learning further improves all predictions. Jointly predicting Fermi level and band gap extends the improvement: DOSForge achieves the best DOS (5.70) and EPDOS (7.84), continuing the trend from 1-task to 2-task to 4-task. Scalar predictions also improve over Single (E_F: 0.229 → 0.217 eV, E_g: 0.189 → 0.171 eV). In contrast, naive 4-task learning degrades spectral predictions compared to 2-task Parallel (DOS: 6.08 → 6.22, EPDOS: 8.75 → 8.89), showing that naive multi-task learning cannot exploit these additional targets effectively.

Parameter-free band gap extraction strengthens physical alignment. When the band gap is predicted with separate parameters, it can disagree with the DOS prediction. To check consistency, we report E_g^{ext}, the band gap extracted from the predicted DOS by identifying zero-density edges. 4-task Parallel achieves E_g MAE of 0.178 eV but E_g^{ext} of 0.240 eV, revealing significant inconsistency. To isolate the effect of parameter-free extraction, we compare DOSForge with a variant using a learned MLP head for band gap (DOSForge[‡] in Table 2). While DOSForge[‡] achieves reasonable consistency, it degrades band gap (0.171 → 0.192 eV), DOS, and EPDOS, suggesting gradient conflict from the learned head. DOSForge tightly aligns band gap prediction with the predicted DOS by construction, achieving the best extracted-gap accuracy (E_g^{ext} = 0.171 eV) while also improving DOS and EPDOS. This shows that parameter-free band gap extraction strengthens physical alignment and mitigates gradient conflict, enabling band gap supervision to benefit rather than compete with spectral prediction.

5 Conclusion

We presented DOSForge, a multi-task architecture for jointly predicting DOS, element-projected DOS, Fermi level, and band gap from crystal structures. DOSForge couples DOS and EPDOS decoding through cross-conditioning and applies parameter-free band gap supervision, encouraging mutual consistency among electronic descriptors. Across benchmarks and ablations, this design yields state-of-the-art DOS prediction while improving EPDOS and tightening agreement between predicted DOS and extracted gaps. In contrast, naive multi-task learning exhibits clear trade-offs,

highlighting the need for structured inter-task coupling. Overall, DOSForge points toward more general electronic-structure models that learn multiple descriptors jointly, rather than treating them as isolated targets. Future work includes extending this framework to orbital/spin-resolved DOS and band-structure prediction.

ACKNOWLEDGEMENTS

This work was supported by the Technology development Program (RS-2025-25462152) funded by the Ministry of SMEs and Startups (MSS, Korea).

REFERENCES

Junwen Bai, Yuanqi Du, Yingheng Wang, Shufeng Kong, John Gregoire, and Carla P Gomes. Xtal2dos: Attention-based crystal to sequence learning for density of states prediction. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

Frederick N Fritsch and Ralph E Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

Shufeng Kong, Francesco Ricci, Dan Guevarra, Jeffrey B Neaton, Carla P Gomes, and John M Gregoire. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nature communications*, 13(1):949, 2022.

Dilfira Kudrat, Zongxia Xie, Yanru Sun, Tianyu Jia, and Qinghua Hu. Patch-wise structural loss for time series forecasting. 2025.

Namkyeong Lee, Heewoong Noh, Sungwon Kim, Dongmin Hyun, Gyoung S Na, and Chanyoung Park. Density of states prediction of crystalline materials via prompt-guided multi-modal transformer. volume 36, pp. 61678–61698, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2017.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Jiayang Xie, Zhihao Zhang, and Xiao-Ming Cao. Density of states prediction by a metric-optimized equivariant graph neural network. *Chemistry of Materials*, 37(16):6313–6322, 2025.

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

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. 2024.

A Model Architecture Details

A.1 Spectral Decoder

The spectral decoder transforms atomic embeddings from the encoder into DOS and EPDOS spectra. Let $H = \{h_i\}_{i=1}^{n} \in \mathbb{R}^{n \times d}$ denote the encoder outputs, where n is the number of atoms and d=256 is the embedding dimension.

Decoding with energy patch tokens. Each spectrum is decoded using N_p=32 learnable energy patch tokens $\mathbf{z} \in \mathbb{R}^{N_p \times d}$, which attend to the atomic embeddings via cross-attention. This cross-attention decoding approach follows prior work on DOS prediction (Bai et al., 2022; Lee et al., 2023). Both DOS and EPDOS branches are initialized with the same learnable tokens. For DOS, all atoms serve as keys and values. For EPDOS, atoms are grouped by element type; each element group ℓ attends only to atoms of that element during cross-attention. All element groups share the same decoder weights; only the atoms they attend to differ.

Transformer block. Each decoder block applies cross-attention, self-attention, and FFN with pre-LayerNorm residual connections:

$$\tilde{\mathbf{z}} = \mathbf{z} + \sigma\left(\frac{\mathrm{LN}(\mathbf{z})\, W_Q\, (\hat{H}\, W_K)^\top}{\sqrt{d_h}}\right) \hat{H}\, W_V, \tag{2}$$

$$\hat{\mathbf{z}} = \tilde{\mathbf{z}} + \mathrm{SelfAttn}\big(\mathrm{LN}(\tilde{\mathbf{z}})\big), \tag{3}$$

$$\mathbf{z}' = \hat{\mathbf{z}} + \mathrm{FFN}\big(\mathrm{LN}(\hat{\mathbf{z}})\big), \tag{4}$$

where $\hat{H} = \mathrm{LN}(H)$ denotes normalized encoder outputs serving as keys and values. Cross-attention uses sigmoid gating (σ) rather than softmax, which avoids probability normalization and can better retain magnitude information; this is useful for EPDOS, where element groups vary in size. Self-attention uses standard softmax.

All attention uses 8 heads with d_h=32; the FFN uses GELU with expansion factor 4. DOS and EPDOS branches use separate LayerNorm modules to normalize the encoder outputs before cross-attention.

Decoding pipeline and intermediate supervision. The decoding pipeline proceeds as follows: L_1=2 blocks $\rightarrow$ unpatchification head (mid predictions) $\rightarrow$ cross-conditioning exchange (Section A.2) $\rightarrow$ L_2=2 branch-specific blocks $\rightarrow$ unpatchification head (final predictions). The intermediate predictions are supervised by intermediate losses $\mathcal{L}^{\mathrm{mid}}$, which encourage meaningful representations before the exchange. We found that these intermediate losses are essential for cross-conditioning to be effective: without them, DOS WD degrades by 2.6% and EPDOS WD by 5.5%. DOS and EPDOS each have their own unpatchification head, which is shared between mid and final predictions within each branch.

Unpatchification head. The unpatchification head maps each patch token to a spectral segment via a per-token MLP: $\hat{\mathbf{d}}_p = \mathrm{MLP}(\mathbf{z}_p) \in \mathbb{R}^{T/N_p}$. The full spectrum is obtained by concatenation: $\hat{\mathbf{d}} = [\hat{\mathbf{d}}_1; \ldots; \hat{\mathbf{d}}_{N_p}] \in \mathbb{R}^T$, where $T/N_p = 256/32 = 8$ bins per patch.

A.2 Cross-Conditioning Exchange

After L_1=2 blocks, intermediate predictions are generated for supervision, and then the two streams exchange representations before continuing with L_2=2 branch-specific blocks. The EPDOS$\rightarrow$DOS direction aggregates element-projected representations via scatter-sum:

$$\mathbf{z}_D^{(\mathrm{init})} = \textstyle\sum_{\ell \in \mathcal{E}(g)} \mathbf{z}_E^{(\ell,\mathrm{mid})}, \tag{5}$$

where $\mathcal{E}(g)$ denotes the set of element groups in graph g. This mirrors $D(E) \approx \sum_\ell D_\ell(E)$: the aggregation directly provides an element-summed representation, so subsequent DOS blocks require only self-attention refinement without further cross-attention.

The DOS→EPDOS direction broadcasts the DOS representation to each element group:

$$\mathbf{z}_E^{(\ell,\text{init})} = \mathbf{z}_D^{(\text{mid})}[g(\ell)]. \tag{6}$$

Since all groups receive the same copy, element-specific specialization is recovered via cross-attention with element-masked atomic embeddings in subsequent branch-specific blocks.

Conditioning variants (Figure 2). All variants use the same transformer block design (Eq. 2). **Parallel**: each branch runs 2 blocks independently; no information exchange. **EPDOS→DOS**: EPDOS branch runs L_1=2 blocks, then its representations are scatter-summed to initialize DOS tokens; DOS branch then runs L_2=2 blocks. **DOS→EPDOS**: DOS branch runs L_1=2 blocks, then its representation is broadcast to each element group to initialize EPDOS tokens; EPDOS branch then runs L_2=2 blocks. **Cross-conditioning**: both branches run L_1=2 blocks, exchange representations (scatter-sum into DOS, broadcast into EPDOS), then each continues with L_2=2 branch-specific blocks. Intermediate losses $\mathcal{L}^{\text{mid}}$ supervise intermediate predictions at the exchange point (cross-conditioning only).

A.3 Fermi Level Prediction

The Fermi level is predicted independently of the spectral decoder via attention-weighted pooling over encoder outputs:

$$\hat{E}_F = \text{MLP}(\textstyle\sum_{i=1}^{n} \alpha_i \, h_i), \quad \alpha_i = \text{softmax}\Big(q_F^\top h_i / \sqrt{d}\Big), \tag{7}$$

where $q_F \in \mathbb{R}^d$ is a learnable query. This operates directly on encoder outputs, as the Fermi level is a global scalar property rather than a spectral quantity.

A.4 Band Gap Extraction Algorithm

The band gap is extracted from the predicted DOS without learnable parameters, providing parameter-free band gap supervision during training (Section 3.4). The spectrum is split at the Fermi level into left (L; decreasing energy) and right (R; increasing energy) halves, both ordered outward from the Fermi level. A soft mask $m_s = \text{ReLU}(1 - k \cdot \mathbf{c}_s)$ weights each bin, where $\mathbf{c}_s = \text{cumsum}(\mathbf{d}_s)$ is a cumulative sum from the Fermi level for $s \in \{L, R\}$. Bins inside the gap (where DOS is near zero) receive weights close to 1, while the weights decrease as cumulative DOS grows, approaching 0 once band states are reached. The total gap width is the sum of masked bins scaled by the energy spacing ΔE. Algorithm 1 summarizes the procedure; all operations are differentiable, enabling gradient flow from the band gap loss through the predicted DOS to the encoder.

Algorithm validation. We applied the extraction to ground-truth DOS from all 62,964 structures and compared against ground-truth band gap values. The resulting MAE is 0.056 eV overall (0.019 eV for metals, 0.120 eV for non-metals), which is smaller than the energy bin width ($\Delta E \approx 0.078$ eV).

Evaluation metric (E_g^{ext}). For fair evaluation, we use a conventional non-differentiable algorithm, different from our differentiable training algorithm, to extract band gap from predicted DOS. Starting from the Fermi level, we scan outward in both directions to find the first bin where DOS exceeds a threshold of 0.02 states/eV. The band edge positions are placed at the boundary between the gap region and the first above-threshold bin. The extracted band gap is the distance between the left and right band edges: $E_g^{\text{ext}} = E_{\text{right}} - E_{\text{left}}$. This metric evaluates how well the predicted DOS reflects the true band gap structure.

B Training Details

Dataset. We use crystal structures from the Materials Project (Jain et al., 2013) with valid DOS calculations. From an initial set of 63,040 structures (41,001 non-magnetic, 21,963 magnetic), we exclude 76 structures containing noble gases (He, Ne, Ar, Kr, Xe) due to insufficient element coverage, yielding 62,964 structures spanning 84 elements. The data is split 7 : 1.5 : 1.5 (train / validation / test), stratified by point group across all 32 crystallographic point groups, resulting in 44,059 / 9,428 / 9,477 structures respectively.

Algorithm 1 Band Gap Extraction from DOS

Require: Fermi-centered DOS spectrum $\mathbf{d} \in \mathbb{R}^T$, sharpness k (k=10), energy spacing $\Delta E \approx 0.078\,\text{eV}$
Ensure: Band gap $\hat{E}_g$
$\mathbf{d} \leftarrow \mathrm{ReLU}(\mathbf{d})$ ▷ Enforce $D(E) \geq 0$
$\mathbf{d}_L \leftarrow \mathrm{flip}(\mathbf{d}_{1:T/2}), \quad \mathbf{d}_R \leftarrow \mathbf{d}_{T/2+1:T}$ ▷ Split at Fermi level, order outward
for $s \in \{L, R\}$ **do**
  $\mathbf{c}_s \leftarrow \mathrm{cumsum}(\mathbf{d}_s)$ ▷ Cumulative sum from Fermi level
  $\mathbf{m}_s \leftarrow \mathrm{ReLU}(1 - k \cdot \mathbf{c}_s)$ ▷ Soft gap mask
end for
$\hat{E}_g \leftarrow \left(\sum_i m_{L,i} + \sum_i m_{R,i}\right) \cdot \Delta E$
return $\hat{E}_g$

Data preprocessing. Raw DOS spectra from the Materials Project use non-uniform energy grids with varying resolution. For spin-polarized calculations, spin-up and spin-down channels are summed to obtain total (spin-degenerate) spectra for both DOS and EPDOS. We align all spectra to the Fermi level and interpolate onto a uniform 256-bin grid over $[-10, 10]$ eV using PCHIP (Fritsch & Carlson, 1980), which avoids oscillatory artifacts and preserves local shape fidelity better than cubic spline interpolation. Normalization statistics are computed from the training data after clipping at the 99.9th percentile to suppress rare extreme outliers. DOS uses a single global mean and standard deviation; EPDOS is normalized per element, with each atomic species having its own mean and standard deviation. Fermi level and band gap are standardized using training-set mean and standard deviation. Training losses are computed in normalized space with clipping applied to suppress outliers. All evaluation metrics are computed on denormalized predictions in original units (states/eV for spectra, eV for scalars).

Optimization. We use AdamW (Loshchilov & Hutter, 2017) with weight decay 10^{-2} and gradient clipping at norm 1.0. The learning rate follows a OneCycleLR schedule (Smith & Topin, 2019): linear warmup from 4×10^{-5} to a peak of 10^{-3} over the first 25% of training (75 epochs), then cosine annealing to 10^{-7}. All models are trained for 300 epochs with a batch size of 64 and bfloat16 mixed precision on a single NVIDIA A6000 GPU.

Loss functions and multi-task balancing. Spectral losses ($\mathcal{L}_D$, $\mathcal{L}_{EP}$) combine MAE with patch-wise structural loss (Kudrat et al., 2025): $\mathcal{L}_{\text{spectral}} = \mathcal{L}_{\text{MAE}} + \lambda_{\text{PS}}\mathcal{L}_{\text{PS}}$, where $\lambda_{\text{PS}} = 0.2$. Scalar losses ($\mathcal{L}_F$, $\mathcal{L}_g$) use MAE. Intermediate losses $\mathcal{L}^{\text{mid}}$ supervise intermediate spectral predictions at the cross-conditioning exchange point (Section A.2), weighted by λ_{mid}=0.5. These are merged into their respective main task losses: $\tilde{\mathcal{L}}_D = \mathcal{L}_D + \lambda_{\text{mid}}\mathcal{L}_D^{\text{mid}}$, $\tilde{\mathcal{L}}_{EP} = \mathcal{L}_{EP} + \lambda_{\text{mid}}\mathcal{L}_{EP}^{\text{mid}}$.

The total loss is a weighted sum of all task losses: $\mathcal{L} = \lambda_D\tilde{\mathcal{L}}_D + \lambda_{EP}\tilde{\mathcal{L}}_{EP} + \lambda_F\mathcal{L}_F + \lambda_g\mathcal{L}_g$, with fixed weights λ_D=λ_{EP}=λ_F=1.0, λ_g=0.05 throughout training.

Model selection. For 2-task spectral experiments (DOS + EPDOS), the best checkpoint is selected by the sum of Wasserstein distances on the validation set: $\text{WD}_{\text{sum}} = \text{WD}_{\text{DOS}} + \text{WD}_{\text{EPDOS}}$. For 4-task experiments, we use the sum of raw (unweighted) losses across all tasks, as Wasserstein distance is not applicable to scalar targets.

Baseline implementations. All baselines (Xtal2DoS, DOSTransformer, EqDOS) are implemented based on publicly available code and trained on the same data splits, preprocessing, and evaluation protocol described above.

C Visualization Examples

To illustrate prediction quality across accuracy levels, we select test-set examples at fixed DOS + EPDOS Wasserstein distance (WD) percentiles from 0th (best) to 95th (worst) in 5% increments. Figure 5 shows predicted and ground-truth spectra for both DOS and EPDOS at each percentile. Each panel displays the material formula, Materials Project ID, and WD percentile in the title. Inset metrics report Wasserstein distance, Fermi level error ($|\Delta E_F|$), and band gap error ($|\Delta E_g|$) for DOS; EPDOS panels show per-element spectra with their aggregate WD.

Figure 5: DOS and EPDOS predictions across Wasserstein distance percentiles on the test set. Each column shows one material at a specific WD percentile (0%–95% in 5% increments). For each material, the top panel shows DOS and the bottom panel shows EPDOS. Solid lines: predictions. Dashed lines: DFT ground truth.