Supplementary Material for Density of States Prediction of Crystalline Mateirals via Prompt-guided Multi-Modal Transformer

A Datasets

In this section, we provide further details on the dataset used for experiments.

A.1 Phonon DOS

We use the **Phonon DOS** dataset following the instructions of the official Github repository[4] of a previous work [9]. This dataset contains 1,522 crystalline materials whose phonon DOS is calculated from density functional perturbation theory (DFPT) by a previous work [36]. Since the provided dataset does not contain crystal system information, we additionally collect the information based on the Materials Project (MP) website [3] on the given each material's unique ID (MP-id).

A.2 Electron DOS

We also use **Electron DOS** dataset that contains 38,889 crystalline materials. The Electron DOS dataset consists of the materials and their electron DOS information that is collected from the MP website [3]. Among the collected data, we exclude the materials that are tagged to include magnetism because the DOS of magnetism materials is not accurate to be directly used for training machine learning models [21]. We consider an energy grid of 201 points ranging from -5 to 5 eV with respect to the band edges with 50 meV intervals and the Fermi energy is all set to 0 eV on this energy grid. Moreover, we normalize the DOS of each material to be in the range between 0 and 1. That is, the maximum and minimum value for each DOS is 1 and 0, respectively, for all materials. Moreover, we smooth the DOS values with the Savitzky-Golay filter with the window size of 17 and polyorder of 1 using scipy library following a previous work [9].

A.3 Data Statistics of Electron DOS dataset in OOD scenarios

As described in the main manuscript, we further evaluate the model performance in two out-of-distribution scenarios: **Scenario 1**: regarding the number of atom species, and **Scenario 2**: regarding the crystal systems. We provide detailed statistics of the number of crystalline materials for each scenario in Table 5 and Table 6.

Table 5: The number of crystals according to the number of atom species (Scenario 1).

	Unary (1)	Binary (2)	Ternary (3)	Quaternary (4)	Quinary (5)	Senary (6)	Septenary (7)	Total
# Materials	386	9,034	21,794	5,612	1,750	279	34	38,889

Table 6: The number of crystals according to different crystal systems (Scenario 2).

	Cubic	Hexagonal	Tetragonal	Trigonal	Orthorhombic	Monoclinic	Triclinic	Total
# Materials	8,385	3,983	5,772	3,964	8,108	6,576	2,101	38,889

[4] `https://github.com/zhantaochen/phonondos_e3nn`

B Evaluation Protocol

Phonon DOS. As described in the main manuscript, we evaluate the model performance based on the data splits given in a previous work [9].

Electron DOS. On the other hand, for the Electron DOS dataset, we use different dataset split strategies for each scenario. For the in-distribution setting, we randomly split the dataset into train/valid/test of 80/10/10%. On the other hand, for the out-of-distribution setting, we split the dataset regarding the structure of the crystals. For both scenarios, we generate training sets with simple crystal structures and a valid/test set with more complex crystal structures, because it is crucial to transfer the knowledge obtained from simple crystal structures to that from complex structures in real-world materials science. More specifically, in the scenario 1 (different number of atom species, i.e., # Atom species in Table 2), we use Binary and Ternary materials as training data and Unary, Quaternary, and Quinary materials as valid and test data. In the case of Unary, we exclude it from training data despite its simplicity due to the observed difficulty of the structure, as will be discussed in Section E.1. In the scenario 2 (different crystal systems, i.e., Crystal System in Table 2), we use Cubic, Hexagonal, Tetragonal, Trigonal, and Orthorhombic crystal systems as training set and Monoclinic and Triclinic as valid and test set. In this scenario, where no prompt is available for unseen crystal systems, we employ the mean-pooled representations of the trained prompts during testing, i.e., for the Monoclinic and Triclinic crystal systems. Please refer to Table 5 and Table 6 for detailed statistics of crystals in each scenario.

Physical Properties. In addition to evaluating the accuracy of the model's predictions of the DOS, it is crucial to assess the physical meaningfulness of the predicted DOS for real-world applications. To assess the physical meaningfulness of the predicted DOS, we utilize the predicted DOS to estimate a range of important material properties. Specifically, we evaluate three materials' properties: the bulk modulus for phonon DOS, and the band gap and Fermi energy for electron DOS (Table 1).

Bulk Modulus [5] is a thermodynamic quantity measuring the resistance of a substance to compression. It provides a measure of the material's ability to withstand changes in volume under applied pressure. In the context of elastic properties, the bulk modulus serves as a descriptor, as it indicates how well a material can recover its original volume after being subjected to compression.

Another property we focus on is the Band Gap [6], which refers to the energy range in a material where no electronic states exist. It represents the energy difference between the top of the valence band and the bottom of the conduction band in insulators and semiconductors. Functional inorganic materials, such as those used in applications like LEDs, transistors, photovoltaics, or scintillators, require a comprehensive understanding of their band gap [62]. By accurately predicting the band gap based on DOS, we can accelerate the development of new materials for a wide range of applications.

Additionally, we predict the Fermi Energy [7], which represents the highest energy level occupied by electrons at absolute zero temperature (0K). It can be used to determine the electrical and thermal characteristics of materials.

C Implementation Details

In this section, we provide implementation details of DOSTransformer.

Graph Neural Networks. Our graph neural networks consist of two parts, i.e., encoder and processor. Encoder learns the initial representation of atoms and bonds, while the processor learns to pass the messages across the crystal structure. More formally, given an atom v_i and the bond e_{ij} between atom v_i and v_j, node encoder ϕ_{node} and edge encoder ϕ_{edge} outputs initial representations of atom v_i and bond e_{ij} as follows:

$$\mathbf{h}_i^0 = \phi_{node}(\mathbf{X}_i), \quad \mathbf{b}_{ij}^0 = \phi_{edge}(\mathbf{B}_{ij}), \tag{4}$$

where $\mathbf{X}$ is the atom feature matrix whose i-th row indicates the input feature of atom v_i, $\mathbf{B} \in \mathbb{R}^{n \times n \times F_e}$ is the bond feature tensor with F_e features for each bond. With the initial representations

[5] https://en.wikipedia.org/wiki/Bulk_modulus

[6] https://en.wikipedia.org/wiki/Band_gap

[7] https://en.wikipedia.org/wiki/Fermi_energy

of atoms and bonds, the processor learns to pass messages across the crystal structure and update atoms and bonds representations as follows:

$$\mathbf{b}_{ij}^{l+1} = \psi_{edge}^{l}(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{b}_{ij}^l), \quad \mathbf{h}_i^{l+1} = \psi_{node}^{l}(\mathbf{h}_i^l, \sum_{j \in \mathcal{N}(i)} \mathbf{b}_{ij}^{l+1}), \tag{5}$$

where $\mathcal{N}(i)$ is the neighboring atoms of atom v_i, ψ is a two-layer MLP with non-linearity, and $l = 0, \ldots, L'$. Note that $\mathbf{h}_i^{L'}$ is equivalent to the i-th row of the atom embedding matrix $\mathbf{H}$ in Equation 1.

Model Training. In all our experiments, we use the AdamW optimizer for model optimization. For all the tasks, we train the model for 1,000 epochs with early stopping applied if the best validation loss does not change for 50 consecutive epochs.

Hyperparameter Tuning. Detailed hyperparameter specifications are given in Table 7. For the hyperparameters in DOSTransformer, we tune them in certain ranges as follows: number of message passing layers in GNN L' in {2, 3, 4}, number of cross-attention layers L_1, L_3 in {2, 3, 4}, number of self-attention layers L_2 in {2, 3, 4}, hidden dimension d in {64, 128, 256}, learning rate η in {0.0001, 0.0005, 0.001}, and batch size B in {1, 4, 8}. We use the sum pooling to obtain the crystalline material i's representation, i.e., $\mathbf{g}_i$. We report the test performance when the performance on the validation set gives the best result.

Table 7: Hyperparameter specifications of DOSTransformer.

Hyperparameters	**In-Distribution**		**Out-of-Distribution**	
	Phonon DOS	**Electron DOS**	**# Atom Species**	**Crystal Systems**
# Message Passing Layers (L')	3	3	3	3
# Cross-Attention Layers (L_1)	2	2	2	2
# Self-Attention Layers (L_2)	2	2	2	2
# Cross-Attention Layers (L_3)	2	2	2	2
Hidden Dim. (d)	256	256	256	256
Learning Rate (η)	0.0001	0.0001	0.0001	0.0001
Batch Size (B)	1	8	8	8

D Methods Compared

In this section, we provide further details on the methods that are compared with DOSTransformer in our experiments.

MLP. We first encode the atoms in a crystalline material with an MLP. Then, we obtain the representation of material i, i.e., $\mathbf{g}_i$, by sum pooling the representations of its constituent atoms. With the material representation, we predict DOS with an MLP predictor ϕ', i.e., $\hat{\mathbf{Y}}^i = \phi'(\mathbf{g}_i)$, where $\phi' : \mathbb{R}^d \rightarrow \mathbb{R}^{201}$.

On the other hand, when we incorporate energy embeddings into the MLP, we predict DOS for each energy j with a learnable energy embedding $\mathbf{E}_j^0$ and obtained material representation $\mathbf{g}_i$, i.e., $\hat{\mathbf{Y}}_j^i = \phi(\mathbf{E}_j^0 || \mathbf{g}_i)$, where $\phi : \mathbb{R}^{2d} \rightarrow \mathbb{R}^1$ is a parameterized MLP.

Graph Network. We first encode the atoms in a crystalline material with a graph network [4]. As done for MLP, we obtain the representation of material i, i.e., $\mathbf{g}_i$, by sum pooling the representations of its constituent atoms. With the material representation, we predict the DOS with an MLP predictor, i.e., $\hat{\mathbf{Y}}^i = \phi'(\mathbf{g}_i)$, where $\phi' : \mathbb{R}^d \rightarrow \mathbb{R}^{201}$. Note that the only difference with MLP is that the atom representations are obtained through the message passing scheme. We also compare the vanilla graph network that incorporates the energy information as we have done in MLP.

E3NN. For E3NN [9], we use the official code published by the authors[8], which implements equivariant neural networks with E3NN python library[9]. By learning the equivariance, the model can generate high-quality representations with a small number of training materials. After obtaining the crystalline material representation $\mathbf{g}_i$, all other procedures have been done in the same manner with other baseline models, i.e., MLP and Graph Network.

E Additional Experiments

E.1 Model Performance Analysis on Out-of-Distribution Scenarios

In this section, we conduct a comprehensive analysis of the model's predictions in the out-of-distribution scenarios presented in Table 2. In Table 8, we evaluate the performance of the model for each type of material, providing detailed insights into its predictive capabilities. We have following observations: **1)** We observe that DOSTransformer consistently outperforms in both out-of-distribution scenarios, which demonstrates the superiority of DOSTransformer. **2)** The performance of all the compared models generally degrades as the crystal structure gets more complex. That is, models perform worse in Quinary crystals than in Quarternary crystals, and worse in Triclinic crystals than in Monoclinic crystals. **3)** On the other hand, it is not the case in Unary crystal. This is because in Unary crystal only one type of atom repeatedly appears in the crystal structure, which cannot give enough information to the model. However, DOSTransformer also makes comparably accurate predictions in the Unary materials by modeling the complex relationship between the atoms and various energy levels.

Table 8: Model performance in Out-of-Distribution scenarios.

Model	**# Atom Species**						**Crystal System**			
	Unary		**Quarternary**		**Quinary**		**Monoclinic**		**Triclinic**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Energy ✗										
MLP	1.457 (0.022)	0.300 (0.004)	0.737 (0.002)	0.186 (0.001)	0.926 (0.001)	0.206 (0.002)	0.747 (0.013)	0.188 (0.001)	0.841 (0.018)	0.202 (0.001)
Graph Network	0.894 (0.049)	0.222 (0.003)	0.551 (0.015)	0.152 (0.003)	0.712 (0.003)	0.177 (0.004)	0.504 (0.011)	0.145 (0.002)	0.582 (0.005)	0.160 (0.002)
E3NN	0.541 (0.022)	0.164 (0.006)	0.491 (0.001)	0.144 (0.001)	0.716 (0.008)	0.177 (0.000)	0.393 (0.004)	0.130 (0.001)	0.510 (0.008)	0.149 (0.001)
Energy ✓										
MLP	0.501 (0.012)	0.170 (0.002)	0.468 (0.002)	0.147 (0.000)	0.638 (0.008)	0.173 (0.002)	0.402 (0.006)	0.138 (0.001)	0.520 (0.011)	0.156 (0.001)
Graph Network	0.461 (0.001)	0.158 (0.010)	0.420 (0.003)	0.134 (0.001)	0.586 (0.008)	0.162 (0.001)	0.370 (0.012)	0.125 (0.002)	0.479 (0.014)	0.143 (0.002)
E3NN	0.496 (0.019)	0.156 (0.002)	0.479 (0.004)	0.145 (0.001)	0.686 (0.009)	0.177 (0.001)	0.385 (0.002)	0.129 (0.001)	0.502 (0.002)	0.148 (0.001)
DOSTransformer	**0.438** (0.007)	**0.145** (0.002)	**0.407** (0.006)	**0.127** (0.001)	**0.575** (0.007)	**0.155** (0.001)	**0.353** (0.004)	**0.119** (0.001)	**0.467** (0.004)	**0.137** (0.002)

E.2 Injecting Crystal System Information to Baseline Methods

In this section, we adopt our prompt-based crystal system information injection procedure to the baseline methods. We examine two approaches for injecting the information: 1) injecting the information into the input atoms (i.e., Position 1), and 2) injecting it before the DOS prediction layer (i.e., Position 2). In Table 9, we have the following observations: **1)** Compared to Table 1, all baseline models benefit from using crystal system information. This demonstrates the importance of utilizing crystal structural systems information, which has been overlooked in previous works. **2)** However, DOSTransformer still outperforms all baseline methods with crystal system information (See DOSTransformer in Table 1), verifying the importance of an elaborate design of crystal system injection procedure. To be more specific, we notice a relatively significant performance gap between

[8] https://github.com/ninarina12/phononDoS_tutorial

[9] https://docs.e3nn.org/en/latest/index.html

DOSTransformer and the best baseline model in Electron DOS, which comprises a broader range of crystalline materials than Phonon DOS. This finding highlights the importance of an intricate crystal system injection procedure when striving to learn the DOS of diverse crystalline materials.

Table 9: Baseline model performance with crystal structural system prompts.

Model	Phonon DOS				Electron DOS			
	Position 1		Position 2		Position 1		Position 2	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Energy ✗								
MLP	0.357 (0.007)	0.116 (0.001)	0.341 (0.002)	0.114 (0.001)	0.751 (0.010)	0.191 (0.000)	0.759 (0.015)	0.192 (0.003)
Graph Network	0.363 (0.008)	0.104 (0.000)	0.343 (0.023)	0.106 (0.001)	0.317 (0.005)	0.112 (0.001)	0.323 (0.008)	0.113 (0.001)
E3NN	0.210 (0.007)	0.077 (0.002)	0.209 (0.010)	0.079 (0.001)	0.296 (0.005)	0.109 (0.001)	0.301 (0.006)	0.110 (0.001)
Energy ✓								
MLP	0.239 (0.003)	0.099 (0.001)	0.228 (0.003)	0.098 (0.001)	0.316 (0.004)	0.123 (0.001)	0.313 (0.002)	0.122 (0.001)
Graph Network	0.209 (0.003)	0.089 (0.001)	0.204 (0.004)	0.087 (0.000)	0.247 (0.001)	0.101 (0.001)	0.245 (0.004)	0.100 (0.002)
E3NN	0.194 (0.004)	0.073 (0.000)	0.190 (0.000)	0.073 (0.001)	0.291 (0.001)	0.109 (0.001)	0.293 (0.002)	0.112 (0.001)

E.3 Qualitative Analysis

In this section, we provide a qualitative analysis of the predicted DOS by mainly comparing it to the DFT-calculated (i.e., Ground Truth) DOS and our main baseline (i.e., E3NN). In Figure 5 (a), which represents the predicted DOS of materials not containing transitional metals, both E3NN and DOSTransformer successfully capture the overall trend of the DOS for several materials (e.g., mp-13063, mp-10931, mp-1009129, and mp-16378). However, DOSTransformer shows a much more precise prediction that closely aligns with the ground truth DOS, providing even more useful information beyond the shape of DOS. For example, peak points represent regions of high density and are likely to be strongly influenced when materials undergo changes in property, and thus represents the probabilistically important energy regions of the materials in the process of material discovery. Notably, our model better captures the peak points in the ground truth DOS compared to E3NN, demonstrating the applicability of DOSTransformer-predicted DOS for real-world material discovery.

On the other hand, Figure 5 (b) shows the DOS prediction for materials containing transition materials. Although DOSTransformer provides more reliable prediction, we observe that the prediction errors of both models get larger compared to the materials that do not contain transition metals shown in Figure 5 (a). This can be attributed to the inherent complexity of physical properties in materials containing transition metals, as discussed in Section 6. Therefore, for our future work, we plan to design expert models in which each expert is responsible for materials with and without transition metals, to achieve more refined and accurate predictions of the DOS. This approach would enable a more comprehensive and elaborate analysis of the DOS in different material compositions.

E.4 Various Training Data Ratio for Fine-Tuning

In this section, we additionally provide experimental results on various ratios of training data for fine-tuning in Table 3. That is, instead of sampling 10% of training data from the test set used in OOD scenarios in Section 5.3, we try various sampling ratios, i.e., 5%, 10%, 15%, and 20%, from the test set. We have the following observations: **1)** We notice a significant performance disparity between the “Only Prompt" and “All" approaches, particularly when the training dataset is limited. This phenomenon can be attributed to the challenge of overfitting when fine-tuning the entire model on a small subset of materials, as discussed in Section 5.3. **2)** In contrast, when the train-

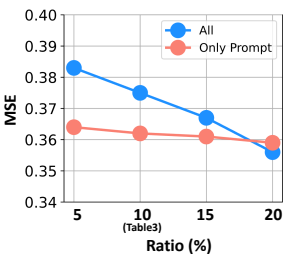

Figure 6: Various training data ratios for fine-tuning.

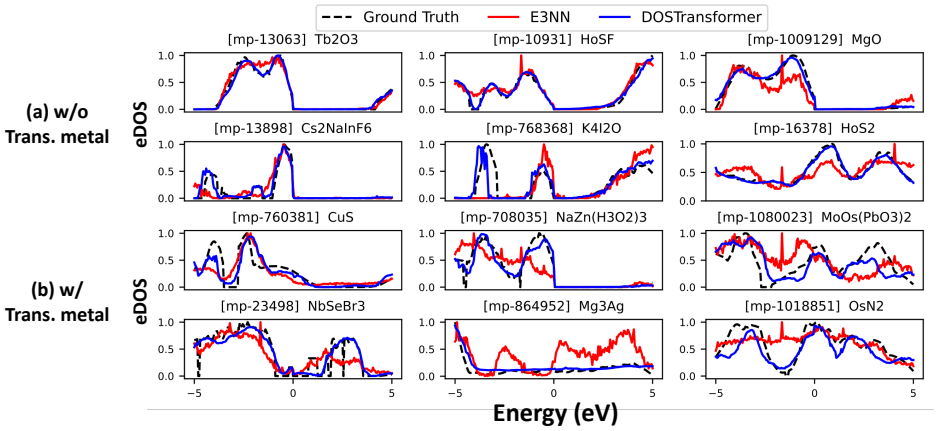

Figure 5: Qualitative Analysis.

ing data becomes more abundant, we find that fine-tuning the entire model parameters surpasses the performance of only tuning the prompt parameters. This observation aligns with the analysis presented in Section 5.3. However, it is important to note that the existing DFT calculation-based databases suffer from a highly biased distribution, which limits their coverage of different materials. This limitation emphasizes the significance of achieving good performance even with a small subset of training data. Therefore, we argue that the application of prompt tuning enhances the real-world applicability of DOSTransformer.

E.5 Model Training and Inference Time

In this section, to verify the efficiency of DOSTransformer, we compare the training and inference time of the baseline methods in Table 10. We observe that DOSTransformer requires a longer training time per epoch on the Phonon DOS dataset compared to E3NN, which can be attributed to the two forward passes (i.e., system and global energy embeddings) during the training procedure. However, when it comes to the Electron DOS dataset, DOSTransformer demonstrates a shorter training time per epoch compared to E3NN. This is because the Electron DOS dataset has complex crystal structures, requiring more time for E3NN to learn equivariant representations. Furthermore, in terms of inference time, DOSTransformer demonstrates significantly faster computation per epoch compared to E3NN, particularly on the Electron DOS dataset. This is because we only utilize system prediction without global prediction during inference. As many predictive ML models are used for high-throughput screening in material discovery, inference time is a critical factor for ML models in materials science, demonstrating the practicality of DOSTransformer in real-world applications.

Table 10: Training and inference time per epoch for each dataset (sec/epoch).

Model	Training		Inference	
	Phonon DOS	Electron DOS	Phonon DOS	Electron DOS
Energy ✗				
MLP	4.10	23.52	1.51	3.00
Graph Network	16.17	59.74	1.95	3.88
E3NN	21.21	141.02	3.72	9.49
Energy ✓				
MLP	4.67	27.88	1.66	3.10
Graph Network	17.45	66.83	2.16	4.28
E3NN	24.12	152.80	3.92	10.21
DOSTransformer	39.17	145.85	2.98	5.99

F Broader Impacts

Potential Positive Scientific Impacts. In this work, we propose DOSTransformer, which is the first work that considers various energy levels during DOS prediction and introduces prompts for crystal structural system, demonstrating its applicability in real-world scenarios. For example, transferring the knowledge obtained from simple structured materials to complex structured materials is crucial because DFT calculation-based databases cover limited types of materials or structural archetypes. Therefore, we believe DOSTransformer has broad impacts on various fields of materials science.

Potential Negative Societal Impacts. This work explores the automation process for materials science without wet lab experiments. However, it is important to acknowledge that in the industry, there are skilled professionals dedicated to conducting such experiments for materials science. Therefore, it is important to proactively address these concerns by encouraging collaboration between automated methods and human experts.

G Pseudo Code

Algorithm 1 shows the pseudocode of DOSTransformer.

Algorithm 1: Pseudocode of DOSTransformer.

Input : An input crystalline material $\mathcal{G} = (\mathbf{X}, \mathbf{A})$, Ground truth DOS $\mathbf{Y}$, Number of attention layers L_1, L_2, L_3, Initialized energy embeddings $\mathbf{E}$, Initialized crystal system prompts $\mathbf{P}$.

$\mathbf{H} \leftarrow \text{GNN}(\mathbf{X}, \mathbf{A})$
$\mathbf{E}^{L_1} \leftarrow \text{Cross-Attention}(\mathbf{H}, \mathbf{E}, L_1)$
$\mathbf{g} \leftarrow \text{Sum Pooling}(\mathbf{H})$

$\mathbf{E}^{glob} \leftarrow (\mathbf{E}^{L_1} || \mathbf{g})$
$\tilde{\mathbf{E}}^{0,glob} \leftarrow \phi_1(\mathbf{E}^{glob})$
$\tilde{\mathbf{E}}^{L_2,glob} \leftarrow \text{Self-Attention}(\tilde{\mathbf{E}}^{0,glob}, L_2)$ `// Global Self-Attention`
$\mathbf{E}^{L_3,glob} \leftarrow \text{Cross-Attention}(\mathbf{H}, \tilde{\mathbf{E}}^{L_2,glob}, L_3)$
$\hat{\mathbf{Y}} \leftarrow \phi_{pred}(\mathbf{E}^{L_3,glob})$
$\mathcal{L}^{glob} \leftarrow \text{RMSE}(\hat{\mathbf{Y}}, \mathbf{Y})$

$\mathbf{E}^{sys} \leftarrow (\mathbf{E}^{L_1} || \mathbf{g} || \mathbf{P})$
$\tilde{\mathbf{E}}^{0,sys} \leftarrow \phi_2(\mathbf{E}^{sys})$
$\tilde{\mathbf{E}}^{L_2,sys} \leftarrow \text{Self-Attention}(\tilde{\mathbf{E}}^{0,sys}, L_2)$ `// System Self-Attention`
$\mathbf{E}^{L_3,sys} \leftarrow \text{Cross-Attention}(\mathbf{H}, \tilde{\mathbf{E}}^{L_2,sys}, L_3)$
$\hat{\mathbf{Y}} \leftarrow \phi_{pred}(\mathbf{E}^{L_3,sys})$
$\mathcal{L}^{sys} \leftarrow \text{RMSE}(\hat{\mathbf{Y}}, \mathbf{Y})$

$\mathcal{L}^{total} \leftarrow \mathcal{L}^{glob} + \beta \cdot \mathcal{L}^{sys}$ `// Calculate total loss`

Function Cross-Attention$(\mathbf{H}, \mathbf{E}^0, L)$**:**
  for $l = 1, 2, \ldots, L$ **do**
    $\mathbf{E}^l \leftarrow \text{Softmax}(\frac{\mathbf{E}^{l-1}\mathbf{H}^\top}{\mathbf{H}})$
  end
  return $\mathbf{E}^L$

Function Self-Attention$(\tilde{\mathbf{E}}^0, L)$**:**
  for $p = 1, 2, \ldots, L$ **do**
    $\tilde{\mathbf{E}}^p \leftarrow \text{Softmax}(\frac{\tilde{\mathbf{E}}^{p-1}\tilde{\mathbf{E}}^{p-1\top}}{\tilde{\mathbf{E}}^{p-1}})$
  end
  return $\tilde{\mathbf{E}}^L$