DeepTracer Diffusion: A Single-Stage Diffusion Model for Accurate Cryo-EM Backbone Segmentation

Nathaniel Jewel[1*] **Haydn Tamura**[1*] **Haowen Guan**[2*] **Dong Si**[1*]
[1]University of Washington Bothell [2]University of Massachusetts Boston

Abstract

Precise atomic-level interpretation of macromolecular structures is vital for understanding biological mechanisms yet remains challenging due to the complex nature of cryo-electron microscopy (cryo-EM) data. Existing approaches have utilized either multiple convolutional neural networks or complex combinations of autoencoder and latent diffusion models to predict atom locations via image segmentation. We introduce DeepTracer Diffusion, a novel framework that leverages a single Denoising Diffusion Probabilistic Model (DDPM) to perform image segmentation, providing higher accuracy in terms of F1 score and predicted residues for predicted backbone atoms.

1 Introduction

Accurate determination of atomic positions and labels in macromolecular structures is crucial to understanding biological functions and processes. Cryo-electron microscopy (cryo-EM) has become an essential tool in structural biology, offering the ability to visualize macromolecules at near-atomic resolution. However, interpreting cryo-EM maps to extract precise atomic models remains a challenging task because of the complexity and variability of the data.

DeepTracer is a deep learning model for protein structure predictions using 4 U-Net models for atoms, backbone, secondary structures, and amino acids to predict atom locations and types (Pfab et al., 2020). Each U-Net learns its respective data type and results in accurate protein structure predictions given a cryo-EM map. DeepTracer performs well only on high-resolution cryo-EM maps. With medium to low-resolution cryo-EM maps, DeepTracer's predictions worsen. The performance loss is due to the U-Nets struggling to discern low resolution data (often containing a lot of noise), and a failure to accurately determine the segmentation for each output.

Diffusion models are a class of generative models that can output new samples of data by iteratively denoising pure Gaussian noise (Ho et al., 2020). These models learn a data distribution by learning to denoise data distorted by random noise. There have been numerous approaches to image segmentation using diffusion, especially in the biomedical sciences. Existing methods use latent diffusion models paired with an autoencoder. The encoder generates latent representations of input data, which are then fed into a latent diffusion model and decoded into the predicted segmentation (Lin et al., 2024). A drawback with an autoencoder and diffusion model setup is the need to run input through multiple models and the potential requirement of training both an autoencoder and a diffusion model.

We present DeepTracer Diffusion. Rather than performing encoding and decoding steps with the diffusion process, we use a single DDPM model. To support direct predictions of atom class labels, we introduce a novel one-hot style reverse diffusion algorithm that produces discrete segmentation masks at each timestep. Our approach achieves higher F1 scores and predicted residues, while not bounded by cryo-EM map resolution and without the need to use encoding/decoding steps.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

2 Related Works

2.1 Diffusion for Image Segmentation

Several approaches have been proposed regarding diffusion probabilistic models for image segmentation task. SegDiff (Amit et al., 2022) utilizes diffusion models to iteratively refine segmentation maps by merging information from input images and current estimations. MedSegDiff (Wu et al., 2022) extends the use of diffusion models to medical imaging, introducing dynamic conditional encoding and a Feature Frequency Parser to enhance segmentation performance across various medical tasks. Furthermore, MedSegDiff was improved in MedSegDiff-V2 (Wu et al., 2023) through the integration of transformer mechanisms.

The versatility of diffusion models is evident in their success in generating a distribution of segmentation masks, demonstrating promising results in medical imaging (Mo et al., 2023b). Similarly, another research effort models panoptic masks using diffusion models, demonstrating competitive performance in both image and video segmentation tasks (Chen et al., 2023).

Latent diffusion models have also demonstrated promising results in image segmentation. A notable project, SDSeg, utilizes a well-trained latent diffusion model for biomedical image segmentation with a single-step reverse process (Lin et al., 2024). In this approach, the predicted noise is used to estimate a latent representation of a segmentation map, which is then passed into a pixel-space decoder. This enables an efficient single-step reverse process. SDSeg boasts competitive inference speeds compared to other segmentation models.

2.2 Protein Structure Modeling From Cryo-EM Maps

Protein structure modeling from cryo-EM maps has seen significant strides. Cryo2Struct (Giri and Cheng, 2024) employs a combination of 3D transformers and Hidden Markov Models (HMM) for de novo modeling of atomic protein structures. This model features a transformer-encoder and a skip-connected decoder for sequence-to-sequence prediction and voxel classification, followed by an HMM to connect predicted atoms and construct protein backbones.

Diffusion techniques have also been used in the modeling of protein structures as demonstrated by several studies. For instance, DiffModeler (Wang et al., 2024) employs a U-net architecture and the Denoising Diffusion Implicit Model (DDIM) framework (Song et al., 2021). In this approach, Gaussian noise is added to the cryo-EM map during the forward diffusion process. In the reverse diffusion process, the model predicts the positions and labels of backbone atoms using Dice Loss.

RFdiffusion (Watson et al., 2023) applies DDPM to generate protein structures from randomly sampled noisy point clouds of atoms. Through iterative denoising steps guided by learned features, RFdiffusion refines the structures, optimizing for specific functional and structural criteria.

3 Method

We propose a novel one-hot style diffusion algorithm (OneHotDiff) for voxel-wise image segmentation. OneHotDiff integrates the iterative denoising of diffusion models with a traditional segmentation model that directly predicts discrete one-hot masks at each timestep. This achieves a direct reverse diffusion process with a one-hot style output.

3.1 Forward Diffusion

Our forward diffusion process, depicted in fig. 1, follows the traditional forward sampling for DDPM (Ho et al., 2020).

In the forward process, Gaussian noise is gradually added over T steps, described by:

$$q(y_{1:T} \mid y_0) = \prod_{t=1}^{T} \mathcal{N}\big(y_t; \sqrt{\alpha_t}\, y_{t-1},\, 1 - \alpha_t \mathbf{I}\big) \tag{1}$$

The noisy sample y_t at timestamp t can be derived from y_0 using the reparameterized equation:

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon \tag{2}$$

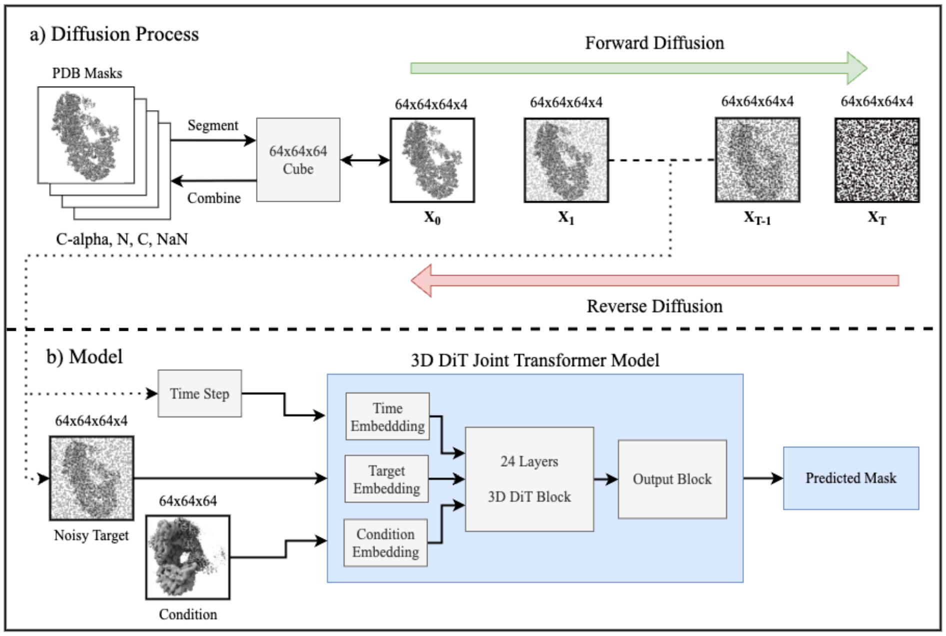

Figure 1: (a) The diffusion process of DeepTracer Diffusion: The process starts with creating masks of classification labels from the PDB files as ground truth. These masks are then segmented into 64x64x64 cubes, each containing a channel for each classification label (Carbon-alpha, Nitrogen, Carbon, and No Atom). The 64x64x64x4 sections are used as the target during the DDIM process. (b) The 3D DiT joint transformer model is trained by randomly sampling a timestep with the pairing of the noised 64x64x64x4 ground truth section with its corresponding 64x64x64 section of the Cryo-EM map.

where $\epsilon \sim \mathcal{N}(0, I)$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\alpha_i \in (0, 1)$ is the diffusion schedule, y_0 is a one-hot mask corresponding to atom classes for each voxel in a 3D grid, and y_t is a weighted combination of the clean one-hot mask and Gaussian noise.

3.2 Reverse Diffusion

The traditional diffusion algorithm reverses the forward process by predicting the noise ϵ in the noisy target y_t. However, image segmentation requires extracting a one-hot prediction $\hat{y}_0$ from y_t, which cannot be obtained by noise prediction alone.

To adapt DDPM for segmentation, we predict the clean one-hot mask $\hat{y}_0$ based on the noisy target y_t and input sample x. Specifically, a segmentation network f_θ estimates the clean mask, and a cold-softmax (temperature $\beta \to 0^+$) enforces a one-hot output:

$$\hat{y}_0 = \mathrm{softmax}_{\beta \to 0^+}\big(f_\theta(x, y_t, t)\big) \tag{3}$$

With $\hat{y}_0$ in hand, we compute the predicted noise:

$$\hat{\epsilon} = \frac{y_t - \sqrt{\bar{\alpha}_t}\, \hat{y}_0}{\sqrt{1 - \bar{\alpha}_t}} \tag{4}$$

For reverse sampling, we replace the stochastic DDPM step with the deterministic DDIM update (Song et al., 2021):

$$y_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{y}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\, \hat{\epsilon} \tag{5}$$

This loop continues until $t = 0$, and the final one-hot mask prediction is $\hat{y}_0$.

4 Evaluation

To evaluate our diffusion-based structure prediction approach, we performed a head-to-head comparison with DeepTracer on a randomly selected benchmark set of 35 cryo-EM maps spanning high to medium resolutions (1.68–5.8 Å, average 4.45 Å). DeepTracer's pipeline is comprised of four specialized 3D convolutional neural networks (CNNs), each implemented as a U-Net architecture and trained for voxel-wise segmentation and coordinate inference (Pfab et al., 2020). The Atoms U-Net classifies each voxel as alpha carbon (C-α), carbon (C), nitrogen (N), or background; the Backbone U-Net labels voxels as backbone, side chain, or non-protein; the Secondary Structure U-Net assigns helix, sheet, or coil conformations; and the Amino Acid Type U-Net predicts one of the twenty standard amino acid identities.

To isolate the impact of our diffusion models, we replaced DeepTracer's Atoms and Backbone U-Net outputs with predictions from two separate diffusion networks: one for atom-type voxel classification (C, C-α, N, background) and one for backbone segmentation. These diffusion outputs were paired with DeepTracer's unchanged Secondary Structure and Amino Acid type U-Nets. All four model outputs were then fed into DeepTracer's standard post-processing and residue-labeling pipeline, which includes oxygen atom placement based on standard backbone geometry. Keeping these downstream steps fixed ensures that any observed performance differences are derive exclusively from the initial voxel classification and coordinate inference stages. The diffusion predictions were generated using a DDIM scheduler with 25 sampling steps to balance inference speed and segmentation fidelity.

Table 1 presents six representative cases from the full test set, illustrating performance at both resolution extremes. Each entry lists EMDB/PDB identifiers, reported resolution, number of deposited residues, predicted residue count, and the resulting F1-score.

Table 1: Six representative cryo-EM maps spanning high to low resolutions, listing EMDB/PDB identifiers, resolution, deposited vs. predicted residue counts, and F1-scores for Diffusion and DeepTracer. The bottom row reports the average resolution, residue counts, and F1-scores over the full 35-map benchmark.

				Diffusion		DeepTracer	
EMDB	PDB	Resolution	Deposited Residues	Predicted Residues	F1-Score	Predicted Residues	F1-Score
emd_20459	6psn	4.60 Å	4154	4101	0.73	1573	0.50
emd_8278	5kp9	5.70 Å	12120	13571	0.63	4814	0.39
emd_3669	5np0	5.70 Å	5056	6810	0.68	978	0.27
emd_46055	9cz0	1.86 Å	2040	2044	0.99	1891	0.96
emd_48671	9mvu	2.20 Å	2105	2121	0.97	1892	0.93
emd_48164	9md1	3.03 Å	764	1058	0.71	867	0.64
Total Average		4.45	3587.60	3817.46	0.75	1798.66	0.59

Across 35 cryo-EM maps, our diffusion-based approach achieved an average F1-score of 0.75 ± 0.0268 (standard error of the mean: SEM) compared to DeepTracer's 0.59 ± 0.0452 (SEM). The resulting 0.16 gap in F1-score far exceeds both SEM value demonstrating a substantial improvement in residues prediction.

We performed a paired t-test on the per-map F1-score differences:

$$t(34) \ = \ 6.86, \quad p \ \approx \ 6.7 \times 10^{-8} \ (< 0.001) \tag{6}$$

which confirms the improvement is highly significant. Moreover, our diffusion networks recover more residues across all resolution ranges, with the largest gains observed in mid-range maps (3–5 Å).

5 Dataset Preparation and Training

5.1 Data Preparation

The dataset is sourced from EMDataResource (Lawson et al., 2016) and consists of 417 cryo-EM maps paired with their corresponding Protein Data Bank (Berman et al., 2000) structures. Among these, 129 maps are high resolution (0–3 Å), 212 are medium resolution (3–5 Å), and 76 are low resolution (>5 Å), spanning an overall resolution range from 2.5 Å to 8.9 Å.

We preprocess the cryo-EM maps by first standardizing the voxel size to 0.5 Å through volume data resampling through UCSF Chimera. This step ensures consistent voxel sizes across all maps, facilitating accurate predictions. Next, we normalize the density values of the maps to a range between 0 and 1.

To create our ground truth, we process the PDB structures using UCSF Chimera to create a mask. We generate a set of masks for each output prediction type. For example: for the atom predictions we make 4 masks, each with a voxel size of 0.5, labeling the voxels as NaN, C-α, C, or N atoms respectively. These masks are then combined via one-hot encoding, assigning each voxel a class label value of 0, 1, 2, or 3.

The one-hot encoded masks are paired with their corresponding cryo-EM map, and both data grids are divided into multiple 64^3 subgrids. We use the inner 50^3 core for predictions, while the outer 7-voxel border is included to enhance border predictions and is ultimately discarded.

To account for class imbalance we compute cross-entropy weights for each label and adjust the cross entropy loss during training. Let $\mathbf{P}$ be a vector of class probabilities, where $\mathbf{P_n}$ is the number of occurrences of class $\mathbf{n}$, $\mathbf{N}$ the total number of classes, and $\mathbf{V}$ the total voxels in the dataset. The class weights $\mathbf{W}$ are calculated as:

$$\mathbf{W} = \frac{\frac{V}{P}}{\frac{1}{N}\sum \frac{V}{P}}, \tag{7}$$

which ensures proper normalization across classes based on voxel distributions.

5.2 Training DDPM for Classification

The training process involves using a classification procedure to calculate the loss used for training. During each training step, we apply the Softmax function to the model's output to obtain probability distributions. Next, we employ the argmax function on the one-hot encoded target to derive the target label. Finally, the cross-entropy loss is computed between the output predictions, the argmax of the target label, and the precalculated cross-entropy weights. The loss used for training is the sum of the cross-entropy loss for every label.

Along with cross-entropy loss, we employ dice loss. Using only cross-entropy loss resulted in excessive voxel predictions. Using dice loss prevents the model from over-predicting voxels by more heavily penalizing false positives. We found that training with both cross-entropy loss and dice loss from scratch leads to unstable gradients, so we warm up our model with a lower learning rate using only cross-entropy loss. After the warmup, when our cross entropy loss is around 0.1, we enable dice loss along with cross entropy.

6 Conclusion

We introduce DeepTracer Diffusion, a novel framework that leverages Denoising Diffusion Implicit Models (DDIM) for direct voxel-wise classification and coordinate inference of backbone atoms from cryo-EM maps. By substituting DeepTracer's Atoms and Backbone U-Nets with two specialized diffusion networks, our approach delivers accurate all-atom structure predictions of protein complexes based solely on their cryo-EM densities. In a head-to-head evaluation on 35 cryo-EM maps, the diffusion model increased the mean F1-score from 0.59 to 0.75 and recovered more residues across the entire resolution range.

In future work, we will scale up our diffusion models and expand the training dataset, reduce the voxel patch size from 4^3 to 2^3 to capture finer structural features, as well as replace DeepTracer's Secondary Structure and Amino Acid Type U-Nets with diffusion-based counterparts. These improvements will push us toward a fully diffusion-driven, end-to-end pipeline for high-fidelity, all-atom model building from cryo-EM maps.

References

T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv:2112.00390 [cs.CV]*, 2022.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. RCSB PDB website: https://www.rcsb.org/; Accessed: 2025-06-06.

Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv:2210.06366*, 2023.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, and Y. Marek. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, 2024.

Nabin Giri and Jianlin Cheng. De novo atomic protein structure modeling for cryoem density maps using 3d transformer and hmm. *Nature Communications*, 15(1):e49647, 2024.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

C. L. Lawson, M. L. Baker, C. Best, C. Bi, M. Dougherty, P. Feng, G. van Ginkel, B. Devkota, S. Patel, G. J. Kleywegt, et al. Emdatabank unified data resource for 3dem. *Nucleic Acids Research*, 44(D1):D396–D403, 2016. EMDataResource statistics page: https://www.emdataresource.org/statistics-historical.html; Accessed: 2025-06-06.

T. Lin, Z. Chen, Z. Yan, W. Yu, and F Zheng. Stable diffusion segmentation for biomedical images with single-step reverse process. *arXiv:2406.18361*, 2024.

S. Mo, E. Xie, R. Chu, L. Yao, L. Hong, M. Nießner, and Z. Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *arXiv:2307.01831*, 2023a.

S. Mo, E. Xie, R. Chu, L. Yao, L. Hong, M. Nießner, and Z. Li. Diffusion models for implicit image segmentation ensembles. *arXiv:2307.01831*, 2023b.

J. Pfab, N. M. Phan, and D. Si. Deeptracer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. *Proceedings of the National Academy of Sciences (PNAS)*, 117(30):17691–17698, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

X. Wang, H. Zhu, G. Terashi, M. Taluja, and D. Kihara. Diffmodeler: Large macromolecular structure modeling in low-resolution cryo-em maps using diffusion model. *bioRxiv*, 2024.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. Vázquez Torres, A. Lauko, V. Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv:2211.00611 [cs.CV]*, 2022.

J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer. *arXiv:2301.11798 [eess.IV]*, 2023.

A Full Test Set Results

Table 2 provides the complete head-to-head performance metrics on all 35 cryo-EM maps in our benchmark. These detailed results extend the representative subset shown in Table 1.

Table 2: Performance comparison between the Diffusion model and DeepTracer on a benchmark test set of 35 cryo-EM maps. Each entry includes the EMDB and PDB identifiers, resolution, and the number of deposited residues. For both methods, we report the number of residues predicted and the resulting F1-score, reflecting residue prediction accuracy. The final row summarizes the average resolution and F1-scores computed across the entire test set.

				Diffusion		DeepTracer	
EMDB	PDB	Resolution	Deposited Residues	Predicted Residues	F1-Score	Predicted Residues	F1-Score
emd_20455	6pqx	4.60	960	1186	0.77	811	0.74
emd_20459	6psn	4.60	4154	4101	0.73	1573	0.50
emd_8278	5kp9	5.70	12120	13571	0.63	4814	0.39
emd_8786	5w9k	4.60	4205	5634	0.69	2720	0.63
emd_8513	5u6r	5.70	14952	7443	0.44	1997	0.20
emd_21136	6vac	5.70	1202	1065	0.55	175	0.21
emd_7439	6ca0	5.75	3698	4362	0.67	781	0.29
emd_3672	5np1	5.70	2460	3087	0.74	1058	0.50
emd_6823	5ydz	5.80	1696	1620	0.71	475	0.38
emd_9577	6kv5	4.60	1679	1944	0.71	865	0.61
emd_9378	6nij	5.70	1972	3211	0.63	520	0.31
emd_8539	5ucy	4.60	2610	2493	0.56	1017	0.37
emd_9541	5gw5	4.60	8446	9319	0.73	5958	0.66
emd_6826	5ye5	5.80	1856	2516	0.69	1291	0.67
emd_3669	5np0	5.70	5056	6810	0.68	978	0.27
emd_8674	5vhf	5.70	6463	6538	0.48	1123	0.22
emd_8735	5vvr	5.80	4565	6390	0.64	582	0.20
emd_6489	3jbw	4.60	1946	2330	0.77	1475	0.70
emd_6906	5zam	5.70	1389	1847	0.61	649	0.47
emd_4400	6i2t	5.70	2252	3044	0.69	555	0.34
emd_5645	3j3x	4.60	8160	9342	0.70	5486	0.65
emd_3790	5oej	5.70	2825	4096	0.59	915	0.39
emd_3963	6evy	4.00	5166	5086	0.96	4370	0.89
emd_3949	6esh	5.10	738	878	0.70	235	0.42
emd_7020	6ayg	4.65	1756	1825	0.81	956	0.66
emd_46055	9cz0	1.86	2040	2044	0.99	1891	0.96
emd_46537	9d3l	2.80	752	743	0.97	708	0.97
emd_70156	9o61	1.68	2040	2117	0.98	2131	0.97
emd_65082	9vib	2.26	566	564	1.00	547	0.98
emd_60984	9iy4	2.00	2646	2597	0.98	2442	0.96
emd_39365	8ykd	2.90	1225	1188	0.98	1116	0.95
emd_48671	9mvu	2.20	2105	2121	0.97	1892	0.93
emd_48164	9md1	3.03	764	1058	0.71	867	0.64
emd_46506	9d30	3.74	2324	2750	0.81	1644	0.76
emd_19930	9es0	2.58	8778	8691	0.99	8336	0.97
Total Average		4.45	3587.60	3817.46	0.75	1798.66	0.59

B Architecture

Our model extends the 3D Diffusion Transformer (DiT) framework of Mo et al. (2023a) by integrating the joint-conditioning transformer design of Stable Diffusion (Esser et al., 2024). As shown in fig. 2, each sub-model (M) has its own multi-head output layer. The current model uses 2 sub-models (Atoms and Backbone) for replacing the Atoms and Backbone U-nets of DeepTracer.

To keep our implementation cleanly compatible with the existing DeepTracer pipeline, we process cubic input volumes of size 64^3 voxels. In early tests, a finer patch embedding size ($p = 2$) yielded sharper features but increased the token count eightfold—exceeding the 48 GB memory of our NVIDIA RTX A6000 GPUs. Consequently, we settled on:

- Transformer layers (d): 24
- Patch embedding: 4
- hidden dimensional: 768
- Attention heads: 16

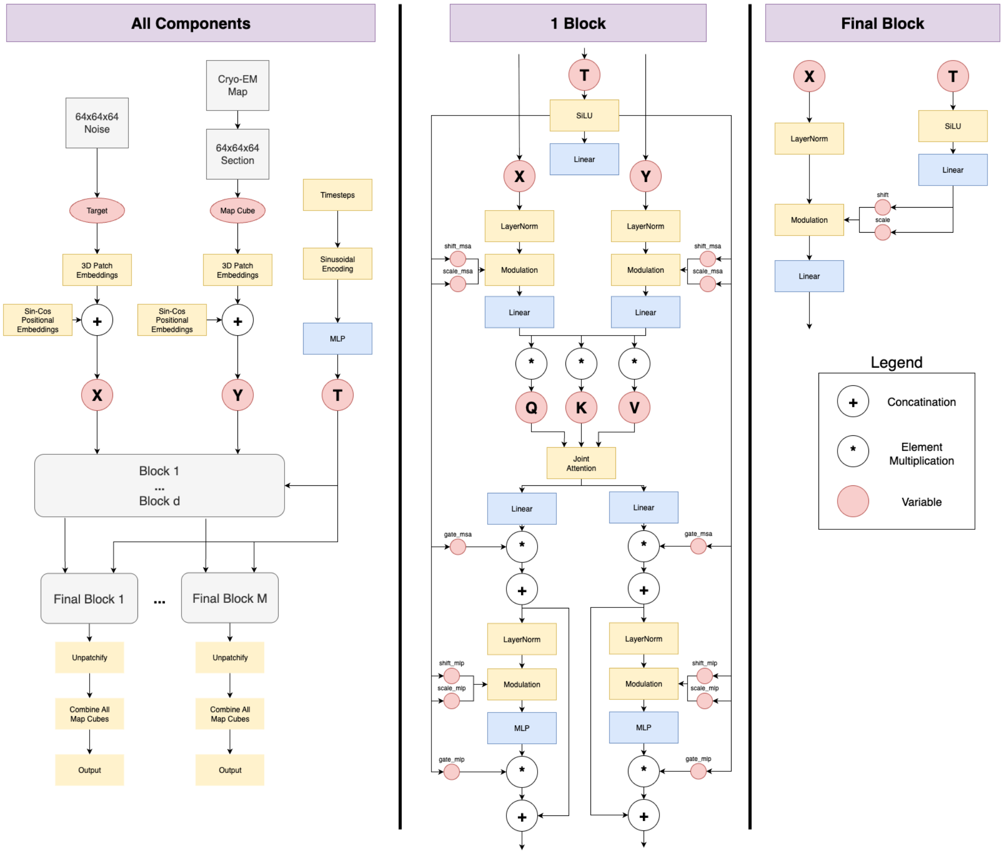

Figure 2: Our model architecture. Showing the full overview, an individual transformer block, and an individual final layer.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that DeepTracer Diffusion replaces U-Net-based segmentation with a diffusion-based generative model, improving performance across cryo-EM resolutions. These claims are supported by benchmark results and detailed methodology in the Evaluation and Dataset Preparation sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the 48 GB GPU memory constraints that led us to select a larger patch embedding size and highlight how this trade-off influences model capacity. We also acknowledge that our evaluation is limited to 35 cryo-EM maps and describe how residue recovery varies across resolution ranges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The focus of the paper is empirical, centered on architectural design, training methodology, and performance evaluation of a diffusion-based segmentation model for cryo-EM data.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper clearly describes the OneHotDiff algorithm, diffusion model training procedure, and evaluation procedure, including diffusion mechanics, segmentation strategy, and dataset pre-processing. These details provide a reproducible framework for validating the main experimental results without requiring access to code or data

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code and trained model weights have not been released. Access to raw data and pre-processing is described.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the resolution-based data split (high, medium, low), voxel resampling procedure, normalization, one-hot mask encoding, and model architecture. These descriptions provide the context for the F1-score differences and residue recovery improvements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Mean F1-scores, standard error of the mean, and a paired t-test are all used in the evaluation section to show statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Justification: We report the GPU memory limit (48 GB on NVIDIA RTX A6000), but omitted run times and total compute.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics. All dataset used are openly licensed, community-curated repositories with clear provenance and no personally identifiable information.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While the paper focuses mainly technical performance / evaluation and we do briefly go over positive societal impact of de-novo cryo-EM atomic structure modeling.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work uses only publicly available, non-sensitive cryo-EM maps and PDB structures.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: We cite EMDataResource and the PDB as data sources. Both are part of the public domain.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The research does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The research does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.