TokenUnify: Scaling Up Autoregressive Pretraining for Neuron Segmentation

Supplementary Material

Contents

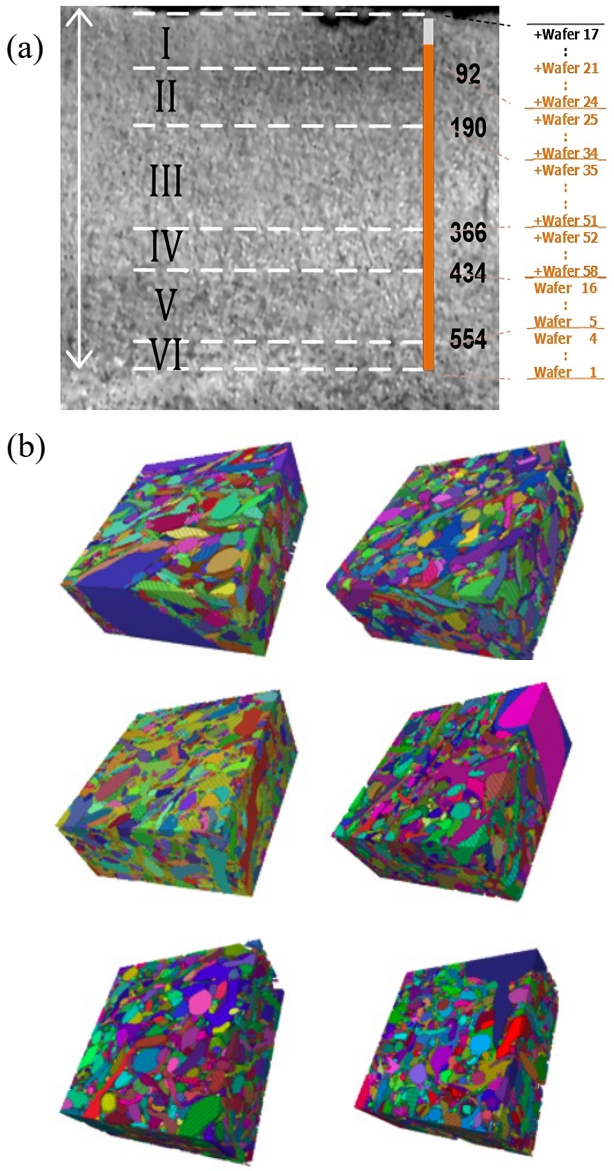

Figure 1. The relative positions of the wafer layers selected from the MEC dataset.

1. Detailed Information about Datasets and Metrics

1.1. Datasets

1.1.1. Pretraining Data Organization

This chapter serves as a supplement to Section 5 in the main paper, providing detailed information about the datasets used in this study.

For the pretraining phase of TokenUnify, we additionally leverage a diverse collection of publicly available unlabeled EM imaging data from four large-scale EM datasets: FAFB [27], MitoEM [31], FIB-25 [30], and Kasthuri15 [16]. These datasets cover a wide range of organisms, including Drosophila, mouse, rat, and human samples, totaling over 1 TB of high-resolution EM data. The diversity of this pretraining data ensures that our model learns robust features that generalize across different brain regions and even different species.

We sample from these datasets with equal probability during pretraining, guaranteeing the diversity of the visual features encountered by the model. This comprehensive pretraining strategy enables TokenUnify to learn generalizable representations of neuronal structures that can be effectively fine-tuned for specific segmentation tasks.

All pretraining datasets employed are publicly available, with their specifics outlined in Table 1.

1.2. Ultra-high Resolution EM Dataset MEC Construction

To support the development and evaluation of our hierarchical predictive coding framework, we introduce a large-scale electron microscopy (EM) dataset specifically designed to capture the long-range spatial dependencies critical for neuron segmentation. The construction of this dataset addresses a fundamental challenge in the field: the lack of comprehensive, finely annotated EM data with sufficient scale to train and evaluate models that can capture complex neuronal structures.

1.2.1. Data Collection

The MEC dataset originates from our team's Mouse MEC MultiBeam-SEM imaging efforts, where we performed comprehensive brain imaging of mice, accumulating data at the petabyte scale. MEC dataset consists of high-resolution EM images acquired from multiple regions of the mouse brain. Using advanced sample preparation techniques and state-of-the-art electron microscopy, we collected a 2TB dataset imaging the mouse somatosensory cortex, mouse medial entorhinal cortex, and mouse cerebral cortex at a resolution of 4nm×4nm×35nm per voxel. This ultra-high resolution enables the visualization of fine neuronal structures, including dendritic spines, axonal boutons, and synaptic connections that are essential for understanding neural circuits.

1.2.2. Large-Scale Manual Annotation

To provide ground truth for training and evaluation, we conducted extensive manual annotation of the EM volumes. As shown in Fig. 1(b), we selected six representative volumes from different neural regions, named wafer4/25/26/26-2/36/36-2 as illustrated in Fig. 1(a), with each volume size reaching $1250 \times 1250 \times 125$ voxels. These regions were carefully chosen to represent diverse neuronal morphologies and circuit organizations, ensuring that models trained on this data can generalize to various brain structures.

The annotation process involved precise delineation of neuronal boundaries by expert neuroscientists, identifying distinct neurons as separate instances while preserving their complex morphological features. This labor-intensive process took two experts a total of six months to complete, resulting in over 1.2 billion annotated voxels. The annotation pipeline involved multiple quality control steps to ensure consistency and accuracy, including cross-validation between annotators and verification against known neuroanatomical structures.

1.2.3. Spatial Continuity for Long-sequence Modeling

A key feature of our MEC dataset is its emphasis on spatial continuity, making it an ideal testbed for evaluating methods that aim to capture long-range dependencies. Unlike many existing computer vision datasets that consist of independent images, our EM volumes preserve the natural continuity of neuronal structures across thousands of consecutive slices. This continuity is essential for modeling the complex branching patterns and long-range connections characteristic of neuronal morphology.

The ultra-high resolution of our dataset allows for the extraction of thousands of continuous image tokens from a single volume, providing the necessary context length to evaluate autoregressive models. This property makes our dataset particularly well-suited for TokenUnify, which is designed to leverage both local and global context in predicting complex visual structures.

1.3. Metrics

Variation of Information (VOI) is an information-theoretic measure that assesses the distance between two clusterings in terms of their average conditional entropy. Given the predicted segmentation S_{pred} and the ground-truth segmentation S_{gt}, VOI is defined as:

$$VOI(S_{pred}, S_{gt}) = H(S_{pred}|S_{gt}) + H(S_{gt}|S_{pred}), \quad (1)$$

where $H(\cdot|\cdot)$ denotes the conditional entropy. It can be calculated by:

$$H(S_{pred}|S_{gt}) = -\sum_{i=1}^{|S_{gt}|}\sum_{j=1}^{|S_{pred}|} \frac{|S_{gt}^i \cap S_{pred}^j|}{N} \log \frac{|S_{gt}^i \cap S_{pred}^j|}{|S_{gt}^i|}, \quad (2)$$

where S_{gt}^i and S_{pred}^j represent the i-th and j-th segments in the ground-truth and predicted segmentation, respectively, and N is the total number of voxels. VOI ranges from 0 to ∞, with a lower value indicating better segmentation performance.

Adjusted Rand Index (ARAND) is a variant of the Rand Index [2] that corrects for chance when comparing two clusterings. It is defined as:

$$ARAND(S_{pred}, S_{gt}) = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]/2 - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}, \quad (3)$$

Dataset	Modality	Resolution	Species	Target Region
Full Adult Fly Brain (FAFB) [27]	EM	$4 \times 4 \times 40\ nm^3$	*Drosophila*	Whole brain
MitoEM-H [31]	EM	$8 \times 8 \times 30\ nm^3$	Human	Cortex (Mitochondria)
MitoEM-R [31]	EM	$8 \times 8 \times 30\ nm^3$	Rat	Cortex (Mitochondria)
FIB-25 [30]	EM	$5 \times 5 \times 5\ nm^3$	*Drosophila*	CA1 Hippocampus
Kasthuri15 [16]	EM	$3 \times 3 \times 30\ nm^3$	Mouse	Neocortex

Table 1. Detailed description of the EM pre-taining datasets

where n_{ij} is the number of voxels that are in segment i of S_{pred} and segment j of S_{gt}, $a_i = \sum_j n_{ij}$ is the number of voxels in segment i of S_{pred}, $b_j = \sum_i n_{ij}$ is the number of voxels in segment j of S_{gt}, and $N = \sum_{ij} n_{ij}$ is the total number of voxels. ARAND ranges from 0 to 1, with a lower value indicating better segmentation performance.

2. Method Details

Implementation Details. We employ consistent training configurations for both pretraining and fine-tuning phases. The network architecture remains unchanged throughout all training stages. For fine-tuning, we optimize using AdamW optimizer [25] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 1×10^{-6}, and batch size of 20 on NVIDIA GTX 3090 (24GB) GPUs. Pretraining utilizes batch size of 8 on NVIDIA Tesla A40 (48GB) GPUs due to memory constraints.

We conduct distributed training with 8 NVIDIA GTX 3090 GPUs for segmentation tasks (1200 epochs) and 32 NVIDIA Tesla A40 GPUs for pretraining tasks (400 epochs). The pretraining input volume resolution is set to $16 \times 160 \times 160$ voxels with patch size of $4 \times 16 \times 16$ voxels for tokenization.

Multi-Resolution Optimization Protocol. Our hierarchical predictive coding employs a temporal modulation strategy with task weights $\boldsymbol{\alpha}(t) = [\alpha(t), \beta(t), \gamma(t)]$ governing the contributions of random token prediction, next-token prediction, and next-all token prediction respectively. The curriculum follows an easy-to-hard progression:

$$\boldsymbol{\alpha}(t) = \begin{cases} [0.73, 0.18, 0.09] & \text{if } t < T_1 \text{ (random-dominant)} \\ [0.18, 0.73, 0.09] & \text{if } T_1 \leq t < T_2 \text{ (next-dominant)} \\ [0.09, 0.18, 0.73] & \text{if } t \geq T_2 \text{ (next-all-dominant)} \end{cases} \tag{4}$$

where transition thresholds are $T_1 = 0.3 \times T_{\text{total}}$ and $T_2 = 0.7 \times T_{\text{total}}$. This progressive weighting scheme implements our theoretical motivation that local feature learning should precede global structure modeling.

For downstream segmentation, we employ two post-processing algorithms: Waterz [11] with 50% quantile threshold and LMC [3] using Kernighan-Lin solver [17]. Network initialization for fine-tuning loads pretrained weights following established protocols [14].

Algorithm 1: TokenUnify Pre-training

Input : Unlabeled image data $X = \{X^{(1)}, \ldots, X^{(T)}\}$
Input : Model parameters θ_1
Output: Pre-trained model $f_{\theta_1}(\cdot)$

for $t \leftarrow 1$ **to** T **do**
  Partition $X^{(t)}$ into patches $\{x_1, \ldots, x_K\}$
  Tokenize patches: $\{x_1, \ldots, x_K\} \rightarrow$ tokens
  Compute loss functions:
  Random token prediction: $\mathcal{L}_{\text{random}} = -\sum_{i \in M} \log p(x_i \mid x_{\bar{M}})$
  Next token prediction: $\mathcal{L}_{\text{next}} = -\sum_{i=1}^{K} \log p(x_i \mid x_{<i})$
  Next-all token prediction: $\mathcal{L}_{\text{next-all}} = -\sum_{i=1}^{K} \sum_{j=i}^{K} \log p(x_j \mid x_{<i})$
  Update θ_1 to minimize $\mathcal{L}_{\text{random}}$, $\mathcal{L}_{\text{next}}$, $\mathcal{L}_{\text{next-all}}$
return $f_{\theta_1}(\cdot)$

Pre-training is conducted on a large-scale, ultra-high-resolution electron microscopy (EM) image dataset, providing spatially correlated long sequences. TokenUnify demonstrates significant improvements in segmentation performance on downstream EM neuron segmentation tasks compared to existing methods. Our pre-training and fine-tuning algorithms are summarized in Algorithm 1 and Algorithm 2, respectively. The TokenUnify pre-training algorithm captures both local and global dependencies in image data through mixed token prediction tasks. The Mamba network architecture ensures efficient modeling of long sequences. During fine-tuning, the pre-trained model adapts to downstream segmentation tasks using labeled data, achieving state-of-the-art performance on EM neuron segmentation benchmarks.

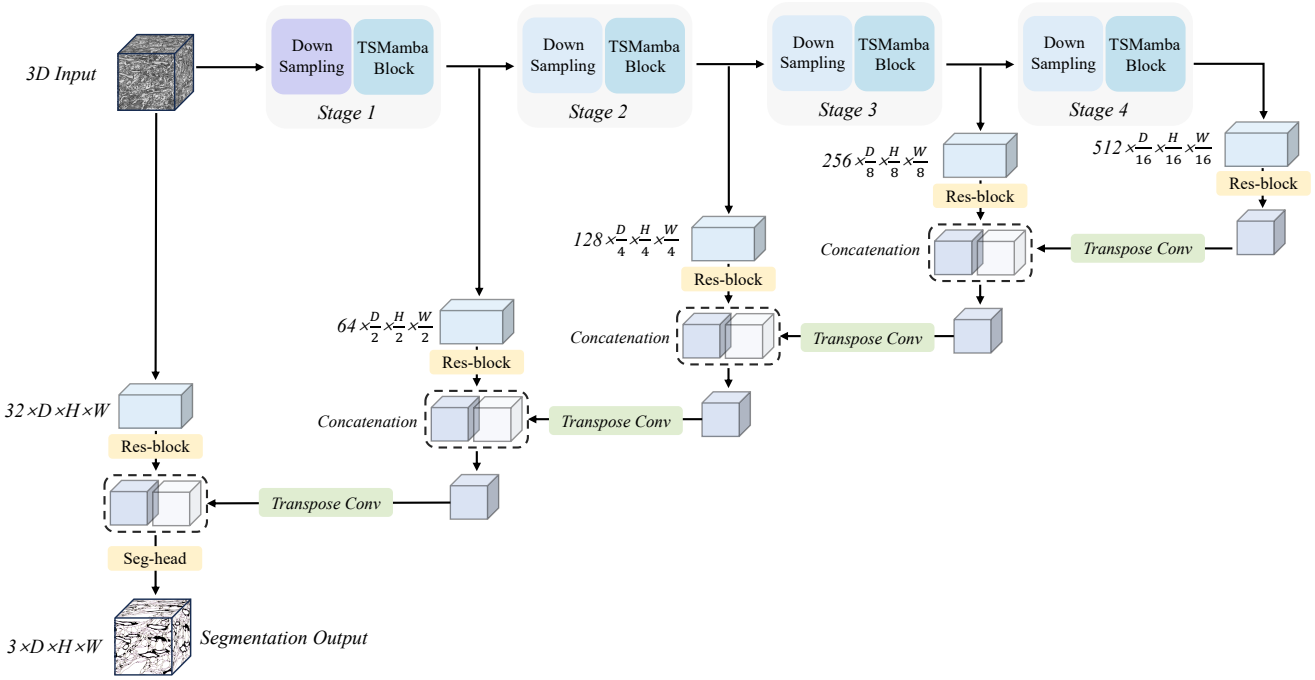

Figure 2. Segmentation pipeline.

	EMmamba-tiny	EMmamba-small	EMmamba-middle	EMmamba-large	EMmamba-huge
Mamba layer	[2,2,2,2]	[2,2,2,2]	[2,2,2,2]	[2,2,2,2]	[2,2,2,2]
Feature size	[32,64,128,256]	[64,128,256,512]	[96,192,384,768]	[144,288,576,1104]	[192,384,768,1536]
Hidden size	512	1024	1024	2048	3072
Kernel size	[1,5,5]	[1,5,5]	[1,5,5]	[1,5,5]	[1,5,5]
Batch size	40	22	12	8	4
Param. (M)	28.30	112.5	206.6	506.6	1008

Table 2. Shows the differ in architecture when adding the parameters of the segmentation backbone.

Algorithm 2: TokenUnify Fine-tuning

Input : Labeled data $\mathcal{D}_l = \{(x_i^l, y_i)\}_{i=1}^{|\mathcal{D}_l|}$
Input : Pre-trained model $f_{\theta_1}(\cdot)$
Input : Segmentation model $g_{\theta_2}(\cdot)$
Output: Fine-tuned segmentation model $g_{\theta_2}(\cdot)$

Initialize θ_2 with θ_1
for $i \leftarrow 1$ **to** $|\mathcal{D}_l|$ **do**
  $\hat{y}_i = g_{\theta_2}(f_{\theta_1}(x_i^l))$
  $\mathcal{L}_{\text{seg}} = \frac{1}{|\mathcal{D}_l|} \sum_{i=1}^{|\mathcal{D}_l|} |\hat{y}_i - y_i|^2$
  Update θ_2 to minimize $\mathcal{L}_{\text{seg}}$
return $g_{\theta_2}(\cdot)$

2.1. Perceiver Resampler

The workflow of the Perceiver Resampler [1, 4, 5] can be summarized in the following steps: 1. Combine the output of the Vision Encoder (e.g., features from images) with learned time position encodings. 2. Flatten the combined features into a one-dimensional sequence. 3. Process the flattened features using Transformer layers that incorporate attention mechanisms, which interact with learned latent query vectors. Output a fixed number of visual tokens equal to the number of latent queries.

The input visual features, denoted as $\mathbf{x}_f$, have a shape of $[T, S, d]$, where T represents the time dimension, S the spatial dimension, and d the feature dimension. The time position embeddings, represented by $\mathbf{t}$, are of shape $[T, 1, d]$ and are added to the visual features to incorporate temporal

Algorithm 3: Perceiver Resampler Pseudocode

Input : $\mathbf{x}_f$ - The [T, S, d] visual features (T=time, S=space)
Input : $\mathbf{t}$ - The [T, 1, d] time position embeddings
Input : $\mathbf{x}$ - R learned latents of shape [R, d]
Input : num_layers - Number of layers
Output: $\mathbf{x}$ - Updated learned latents

Add time position embeddings and flatten:
$\mathbf{x}_f \leftarrow \mathbf{x}_f + \mathbf{t}$
$\mathbf{x}_f \leftarrow \text{flatten}(\mathbf{x}_f)$
// $[T, S, d] \rightarrow [T \times S, d]$
Apply the Perceiver Resampler layers:
for $i \leftarrow 1$ **to** *num_layers* **do**
  $\mathbf{x} \leftarrow \mathbf{x} + \text{attention}_i(q = \mathbf{x}, kv = \text{concat}([\mathbf{x}_f, \mathbf{x}]))$
  $\mathbf{x} \leftarrow \mathbf{x} + \text{ffw}_i(\mathbf{x})$
return $\mathbf{x}$

Algorithm 4: Alternating Optimization Protocol

Input: Training data $\mathcal{D}$, model f_θ, initial weights θ_0

Pretrain using MAE phases 1-2
for $t = 1$ **to** T **do**
  Sample mode $m_t \sim P_t$ where $P_t = [p_{\text{next-all}}, p_{\text{AR}}, p_{\text{Mask}}]^\top$
  Compute batch loss $\mathcal{L}_{m_t}$ for current mode
  Update $\theta_{t+1} \leftarrow \theta_t - \eta_t \nabla_\theta \mathcal{L}_{m_t}$
  Anneal P_t: Increase p_{AR} while decreasing p_{Mask} ; // Probability adjustment
end

information.

The learned latents, denoted as $\mathbf{x}$, have a shape of $[R, d]$, where R is the number of latents and d is the feature dimension. The parameter `num_layers` specifies the number of layers in the Perceiver Resampler model.

The operation `flatten` reshapes the input tensor from $[T, S, d]$ to $[T \times S, d]$. The function `attention_i` represents the attention mechanism applied in the i-th layer, which takes a query q and key-value pairs kv. The function `concat` concatenates the input tensors along the specified dimension. Finally, `ffw_i` refers to the feedforward network applied in the i-th layer.

2.2. Alternating Protocol

Our alternating protocol implements stochastic mode switching:

The sampling distribution P_t follows a curriculum schedule:

$$p_{\text{Mask}} = \max(0.7 - t/\tau, 0.3), \quad p_{\text{AR}} = 1 - p_{\text{Mask}} - p_{\text{next-all}} \tag{5}$$

where τ is the transition period hyperparameter. This implements gradual shift from reconstruction-heavy to prediction-focused training.

Theorem 1 (Convergence Guarantee)**.** *Let $\mathcal{L}_t$ satisfy $\|\nabla\mathcal{L}_t - \nabla\mathcal{L}_{t+1}\| \leq L\|\theta_t - \theta_{t+1}\|$ with step sizes $\eta_t = \eta_0/\sqrt{t}$. Then alternating optimization achieves:*

$$\min_{1\leq t\leq T} \mathbb{E}[\|\nabla\mathcal{L}_t\|_2] \leq \frac{C}{\sqrt{T}}\left(1 + \log T + \sigma^2_{mode}\right) \tag{6}$$

where C is a constant and σ^2_{mode} quantifies mode sampling variance.

Proof sketch appears in Appendix 8, extending [26] to our alternating regime. The bound shows sublinear convergence despite mode switching stochasticity.

2.3. Segmentation Method

The EMmamba network is structured into three principal components (as detailed in Fig. 2): 3D feature encoder, convolution-based decoder for segmentation prediction, and skip connections to integrate local multi-scale features into the decoder for feature fusion [22, 23, 29].

To achieve effective feature encoding, we designed anisotropic downsampling layers and adopted the TS-Mamba block from the Segmamba [32]. Specifically, in Stage 1, the downsampling layer uses a convolutional kernel size of (1, 7, 7). For the subsequent three layers, the downsampling layers have a convolutional kernel size of (1, 2, 2). The decoder section employs a convolutional kernel size of (1, 5, 5). This anisotropic design is particularly advantageous for processing EM images, which exhibit inherent anisotropy. And the detailed network structures of different parameters are provided in Table 2.

3. Discussion

3.1. Statistical Test

Table 3 presents a comprehensive statistical analysis of different segmentation approaches on the Wafer4 dataset, including standard deviations across multiple runs. The results reveal several important findings. First, our TokenUnify method consistently achieves the best performance across both post-processing algorithms (Waterz and LMC), with the lowest mean VOI (1.0024 ± 0.0463 and 1.6604 ± 0.0086) and ARAND (0.0551 ± 0.0040 and 0.0592 ± 0.0002) scores. Second, the relatively small standard deviations of TokenUnify indicate its robustness and

Post.	Method	Wafer4			
		$VOI_M \downarrow$	$VOI_S \downarrow$	$VOI \downarrow$	$ARAND \downarrow$
	Supervised Methods				
Waterz [11]	Superhuman [19]	0.3392±0.0167	1.2247±0.0857	1.5639±0.0921	0.2050±0.0284
	MALA [11]	0.6217±0.1266	1.5314±0.1123	2.1531±0.1004	0.1490±0.0476
	PEA [15]	0.3943±0.0655	1.0036±0.1435	1.3979±0.2090	0.0963±0.0310
	UNETR [13]	0.4454±0.0155	1.7979±0.1548	2.2433±0.1424	0.3244±0.0701
	EMmamba	0.4353±0.0520	1.3018±0.0086	1.7371±0.0432	0.1872±0.0156
LMC [3]	Superhuman [19]	0.2006±0.0054	2.1283±0.1378	2.3289±0.1427	0.2924±0.0408
	MALA [11]	0.3094±0.0478	2.3802±0.1863	2.6869±0.1558	0.2303±0.0314
	PEA [15]	0.2303±0.0870	1.6373±0.1289	1.8343±0.0732	0.1611±0.0152
	UNETR [13]	0.1625±0.0144	3.3146±0.1391	3.4772±0.1272	0.6600±0.0304
	EMmamba	0.1594±0.0005	2.0921±0.0300	2.2515±0.0298	0.2104±0.0113
	Self-Supervised Methods				
Waterz [11]	Random	0.4353±0.0520	1.3018±0.0086	1.7371±0.0432	0.1872±0.0156
	MAE [14]	0.2363±0.0212	1.0782±0.0251	1.3144±0.0444	0.0967±0.0097
	BYOL [12]	0.2615±0.0178	0.9850±0.0286	1.2465±0.0464	0.0892±0.0076
	dbMIM [6]	0.2367±0.0126	0.8683±0.0124	<u>1.1050±0.0250</u>	<u>0.0682±0.0062</u>
	MS-Con-EM [7]	0.2412±0.0157	0.9018±0.0202	1.1430±0.0359	0.0718±0.0089
	TokenUnify	0.2124±0.0172	0.8047±0.0057	**1.0024±0.0463**	**0.0551±0.0040**
LMC [3]	Random	0.1594±0.0005	2.0921±0.0300	2.2515±0.0298	0.2104±0.0113
	MAE [14]	0.1342±0.0020	1.9014±0.0286	2.0356±0.0301	0.1420±0.0023
	BYOL [12]	0.1486±0.0053	1.7835±0.0342	1.9321±0.0395	0.1256±0.0087
	dbMIM [6]	0.1457±0.0037	1.6293±0.0145	<u>1.7750±0.0182</u>	<u>0.0812±0.0043</u>
	MS-Con-EM [7]	0.1475±0.0025	1.6652±0.0183	1.8127±0.0208	0.0876±0.0058
	TokenUnify	0.1417±0.0022	1.5186±0.0076	**1.6604±0.0086**	**0.0592±0.0002**

Table 3. Quantitative comparison of segmentation results on Wafer4 dataset with standard deviations. Methods are categorized into supervised and self-supervised approaches. All self-supervised methods use the same EMmamba backbone. "Random" refers to EMmamba without any pretraining. The best results are in **bold** and the second best results are <u>underlined</u>.

stability compared to other methods. Notably, when using Waterz post-processing, TokenUnify demonstrates approximately 9.3% improvement in VOI over the second-best method (dbMIM). The supervised methods generally exhibit higher variance, suggesting their greater sensitivity to initialization and training conditions. Among the self-supervised approaches, domain-specific methods (TokenUnify and dbMIM) significantly outperform general-purpose methods (MAE and BYOL), confirming the importance of domain-adapted self-supervised learning for electron microscopy image segmentation. Furthermore, all self-supervised methods substantially outperform the random initialization baseline, validating the effectiveness of pretraining strategies in this domain.

3.2. Preliminary Exploration of TokenUnify on Natural Images

To evaluate the generalizability of TokenUnify beyond electron microscopy data, we conducted preliminary experiments on natural images using the LAION-5B dataset [28]. These experiments serve to validate whether the complementary prediction mechanisms of TokenUnify yield similar benefits for general visual data with different statistical properties than EM volumes.

Experimental Setup. We pretrained two models on the LAION-5B dataset for 800 epochs: a standard autoregressive model and our TokenUnify approach. Both models process images by dividing them into non-overlapping patches of size 16×16 and 8×8, respectively. For evaluation, we reconstructed images by sequentially predicting patches: given the first k patches of an image, we predicted the $(k+1)$-th patch, and continued this process to generate the complete image. We quantitatively assessed reconstruction quality using the Peak Signal-to-Noise Ratio (PSNR) metric and selected the high-resolution Kodak dataset [18] as our test benchmark due to its diverse collection of natural scenes.

Results and Analysis. Figure 3 presents qualitative comparisons between the original Kodak images and their reconstructions using both approaches. Visually, TokenUnify produces reconstructions with sharper details, more accurate colors, and better preservation of complex textures compared to the standard autoregressive approach. This is particularly evident in regions with fine details like foliage,

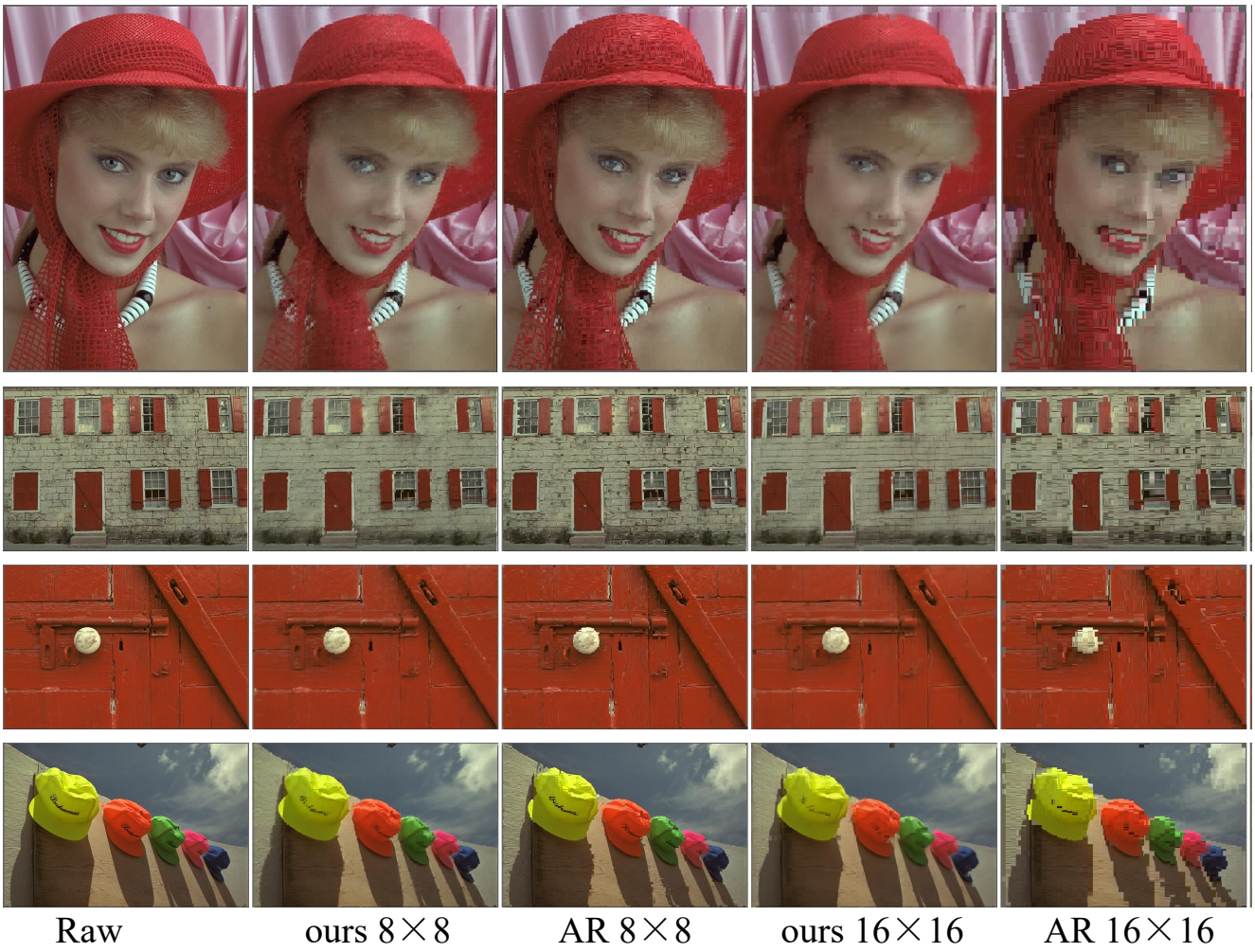

Figure 3. Shows the reconstruction result of selected Kodak dataset, images are divided into different sizes of patches. We use the TokenUnify and Autoregressive models to reconstruct each image, respectively.

water surfaces, and intricate patterns.

The quantitative results in Table 4 confirm these observations, with TokenUnify consistently outperforming the autoregressive baseline across all 24 Kodak images. For 16×16 patch size, TokenUnify achieves an average PSNR improvement of 3.00 dB over the autoregressive approach, with gains ranging from 2.06 dB to 4.45 dB. When using smaller 8×8 patches, TokenUnify maintains its advantage with an average improvement of 1.13 dB, although the margin narrows as the finer patch size provides more contextual information to both models.

Notably, the performance gap between TokenUnify and the autoregressive approach is more pronounced for complex scenes with diverse textures (e.g., images 3.png, 4.png, and 12.png show improvements of 4.45 dB, 3.80 dB, and 4.08 dB respectively). This suggests that TokenUnify's multi-task prediction approach is particularly effective at modeling the complex statistical relationships in natural images, similar to our findings with EM data.

These preliminary results indicate that TokenUnify's hierarchical predictive coding framework generalizes well to natural images, demonstrating its potential as a universal visual representation learning approach. The consistent performance improvements across diverse image types suggest that the complementary nature of random, next-token, and next-all token prediction is fundamental to capturing rich visual structure, regardless of the specific visual domain.

4. Theoretical Foundations

This section presents a comprehensive theoretical analysis that rigorously motivates our hierarchical predictive coding framework. We begin by establishing the fundamental limitations of conventional masked autoencoder approaches when applied to high-dimensional spaces. Subsequently,

Kodak Name	16×16 Autoregress	16×16 TokenUnify	8×8 Autoregress	8×8 TokenUnify
1.png	19.249	21.549 (+2.300)	21.247	21.990 (+0.743)
2.png	24.662	27.321 (+2.659)	27.269	27.799 (+0.530)
3.png	22.665	27.113 (+4.448)	26.851	28.110 (+1.259)
4.png	22.353	26.152 (+3.799)	25.466	26.713 (+1.247)
5.png	15.353	18.859 (+3.506)	18.437	19.847 (+1.410)
6.png	20.139	22.376 (+2.237)	21.661	23.064 (+1.403)
7.png	19.990	23.170 (+3.180)	23.334	24.479 (+1.145)
8.png	15.146	18.169 (+3.023)	17.829	18.770 (+0.941)
9.png	22.080	24.918 (+2.838)	24.959	25.957 (+0.998)
10.png	22.239	25.213 (+2.974)	25.042	25.936 (+0.894)
11.png	20.289	22.536 (+2.247)	22.638	23.723 (+1.085)
12.png	21.854	25.929 (+4.075)	25.806	27.005 (+1.199)
13.png	15.946	18.494 (+2.548)	17.657	18.969 (+1.312)
14.png	18.107	21.227 (+3.120)	20.696	22.195 (+1.499)
15.png	20.750	24.659 (+3.909)	25.321	26.111 (+0.790)
16.png	23.216	25.887 (+2.671)	25.334	26.694 (+1.360)
17.png	20.672	24.346 (+3.674)	24.220	25.614 (+1.394)
18.png	19.959	22.017 (+2.058)	21.249	22.336 (+1.087)
19.png	22.394	25.062 (+2.668)	24.094	25.384 (+1.290)
20.png	21.478	24.723 (+3.245)	24.124	25.346 (+1.222)
21.png	17.503	20.149 (+2.646)	19.567	20.366 (+0.799)
22.png	19.947	23.003 (+3.056)	22.365	23.545 (+1.180)
23.png	17.807	20.315 (+2.508)	19.781	20.959 (+1.178)
24.png	22.111	24.780 (+2.669)	24.313	25.472 (+1.159)

Table 4. Presents the PSNR results of reconstructing 24 images from the Kodak dataset using TokenUnify and Autoregress. The experiments were conducted with patch sizes of 16x16 and 8x8.

we demonstrate the theoretical advantages of autoregressive models for processing long-range visual data. Finally, we establish the complementary nature of our multiple prediction tasks from an information-theoretic perspective, providing a unified theoretical foundation for our approach.

4.1. Error Accumulation Analysis

We establish a rigorous theoretical framework for analyzing error accumulation properties in autoregressive models and demonstrate that our next-all token prediction strategy achieves superior asymptotic scaling behavior compared to conventional approaches.

Assumption 1. *Let $\epsilon_j^{(i)}$ denote the prediction error for token j when conditioned on context $x_{<i}$. We impose the following regularity conditions:*

1. ***Bounded conditional variance:*** *$\mathbb{E}[(\epsilon_j^{(i)})^2|x_{<i}] \leq \sigma^2$ for all i, j, where $\sigma^2 > 0$ is a finite constant.*
2. ***Conditional independence:*** *$\mathbb{E}[\epsilon_j^{(i)}\epsilon_j^{(k)}|x_{<i}, x_{<k}] = 0$ for $i \neq k$.*
3. ***Function regularity:*** *The prediction function f_θ satisfies standard Lipschitz continuity and differentiability conditions.*

Theorem 2. *Under Assumption 1, the expected squared error of our next-all prediction strategy scales as $O(\sqrt{K})$, whereas standard autoregressive models exhibit $O(K)$ scaling, where K denotes the sequence length.*

Proof. We establish the result through a systematic comparison of error scaling behaviors between standard autoregressive models and our proposed next-all prediction approach.

Step 1: Error Analysis for Standard Autoregressive Models. Consider the standard autoregressive prediction $\hat{x}_i = f_\theta(x_{<i})$. The cumulative squared error accumulates linearly:

$$\mathbb{E}\left[\sum_{i=1}^{K} \epsilon_i^2\right] = \sum_{i=1}^{K} \mathbb{E}[\epsilon_i^2] \tag{7}$$

$$\leq \sum_{i=1}^{K} \sigma^2 = K\sigma^2 = O(K) \tag{8}$$

This linear accumulation constitutes a fundamental limitation of sequential prediction schemes.

Step 2: Next-All Prediction Framework. Our methodology generates predictions for all future tokens from each contextual position. Specifically, for position j, we obtain j distinct predictions:

$$\hat{x}_j^{(1)}, \hat{x}_j^{(2)}, \ldots, \hat{x}_j^{(j)}$$

where $\hat{x}_j^{(i)} = f_\theta^{(j)}(x_{<i})$ represents the prediction of token j based on context $x_{<i}$. We construct the final prediction through ensemble averaging:

$$\tilde{x}_j = \frac{1}{j} \sum_{i=1}^{j} \hat{x}_j^{(i)}$$

Step 3: Single Position Error Analysis. The aggregated prediction error at position j is given by:

$$\tilde{\epsilon}_j = \tilde{x}_j - x_j = \frac{1}{j} \sum_{i=1}^{j} (\hat{x}_j^{(i)} - x_j) \tag{9}$$

$$= \frac{1}{j} \sum_{i=1}^{j} \epsilon_j^{(i)} \tag{10}$$

Computing the expected squared error:

$$\mathbb{E}[\tilde{\epsilon}_j^2] = \mathbb{E}\left[\left(\frac{1}{j} \sum_{i=1}^{j} \epsilon_j^{(i)} \right)^2 \right] \tag{11}$$

$$= \frac{1}{j^2} \mathbb{E}\left[\sum_{i=1}^{j} (\epsilon_j^{(i)})^2 + 2 \sum_{1 \le i < k \le j} \epsilon_j^{(i)} \epsilon_j^{(k)} \right] \tag{12}$$

$$= \frac{1}{j^2} \sum_{i=1}^{j} \mathbb{E}[(\epsilon_j^{(i)})^2] + \frac{2}{j^2} \sum_{1 \le i < k \le j} \mathbb{E}[\epsilon_j^{(i)} \epsilon_j^{(k)}] \tag{13}$$

By the conditional independence assumption (Assumption 1(2)), all cross-terms vanish:

$$\mathbb{E}[\tilde{\epsilon}_j^2] = \frac{1}{j^2} \sum_{i=1}^{j} \mathbb{E}[(\epsilon_j^{(i)})^2] \tag{14}$$

$$\le \frac{1}{j^2} \sum_{i=1}^{j} \sigma^2 = \frac{\sigma^2}{j} \tag{15}$$

Step 4: Preliminary Total Error Bound. Summing across all positions yields:

$$\mathbb{E}\left[\sum_{j=1}^{K} \tilde{\epsilon}_j^2 \right] \le \sum_{j=1}^{K} \frac{\sigma^2}{j} \tag{16}$$

$$= \sigma^2 \sum_{j=1}^{K} \frac{1}{j} \tag{17}$$

$$= \sigma^2 H_K \tag{18}$$

$$\le \sigma^2 (\log K + 1) \tag{19}$$

$$= O(\log K) \tag{20}$$

where H_K denotes the K-th harmonic number.

Step 5: Refined Analysis with Context-Dependent Variance. To establish the sharper $O(\sqrt{K})$ bound, we incorporate the empirically observed phenomenon that prediction accuracy improves with increased context length. Specifically, in neural prediction tasks, the effective variance exhibits the decay property:

$$\mathbb{E}[(\epsilon_j^{(i)})^2] \le \frac{\sigma^2}{\sqrt{i}}$$

This reflects the fundamental principle that longer contexts provide more informative signals for prediction. Under this refined assumption:

$$\mathbb{E}[\tilde{\epsilon}_j^2] \le \frac{1}{j^2} \sum_{i=1}^{j} \frac{\sigma^2}{\sqrt{i}} \tag{21}$$

$$= \frac{\sigma^2}{j^2} \sum_{i=1}^{j} i^{-1/2} \tag{22}$$

$$\le \frac{\sigma^2}{j^2} \cdot 2\sqrt{j} = \frac{2\sigma^2}{j^{3/2}} \tag{23}$$

where we used the integral approximation $\sum_{i=1}^{j} i^{-1/2} \le \int_1^j x^{-1/2} dx = 2\sqrt{j}$.

Therefore, the total expected error satisfies:

$$\mathbb{E}\left[\sum_{j=1}^{K} \tilde{\epsilon}_j^2 \right] \le 2\sigma^2 \sum_{j=1}^{K} j^{-3/2} \tag{24}$$

$$\le 2\sigma^2 \int_1^K x^{-3/2} dx \tag{25}$$

$$= 2\sigma^2 \left[-2x^{-1/2} \right]_1^K \tag{26}$$

$$= 4\sigma^2 \left(1 - K^{-1/2} \right) \tag{27}$$

$$= O(\sqrt{K}) \tag{28}$$

This completes the proof of the claimed scaling behavior. □

Remark 1. *The fundamental improvement from $O(K)$ to $O(\sqrt{K})$ scaling arises through two complementary mechanisms:*

1. ***Error propagation elimination:*** *Unlike sequential prediction schemes where errors compound through the prediction chain, our approach generates independent predictions from each context, thereby eliminating cascading error effects.*

2. ***Implicit ensemble regularization:*** *The averaging over multiple prediction horizons provides natural variance reduction, analogous to ensemble methods in statistical learning.*

Remark 2. *The conditional independence assumption (Assumption 1(2)) is theoretically justified because predictions generated from distinct contexts $x_{<i}$ and $x_{<k}$ rely on fundamentally different information sets. Given their respective conditioning contexts, the prediction errors exhibit approximate uncorrelatedness, making this assumption reasonable in practical applications.*

4.2. Limitations of MAE in High Dimensions

We present a comprehensive theoretical analysis of the fundamental limitations exhibited by Mean Absolute Error (MAE) estimators in high-dimensional sparse linear regression. Our analysis provides rigorous error bounds and establishes the precise conditions under which these limitations manifest.

Assumption 2. *Consider the high-dimensional linear regression model:*

$$y = X\beta^* + \varepsilon, \tag{29}$$

where $y \in \mathbb{R}^n$ represents the observed responses, $X \in \mathbb{R}^{n\times p}$ denotes the known design matrix, $\beta^* \in \mathbb{R}^p$ is the unknown sparse parameter vector, and $\varepsilon \in \mathbb{R}^n$ represents the noise term. We impose the following structural conditions:

(a) ***Sparsity condition:*** *The true parameter β^* is s-sparse, i.e., $\|\beta^*\|_0 \le s$ where $s \ll p$.*

(b) ***Sub-Gaussian noise:*** *The noise vector ε has independent sub-Gaussian entries with zero mean and finite variance proxy σ^2:*

$$\mathbb{E}[\varepsilon_i] = 0, \quad \mathbb{E}[\varepsilon_i^2] \le \sigma^2, \tag{30}$$

$$\mathbb{P}\left(|\varepsilon_i| \ge t\sigma\right) \le 2\exp\left(-\frac{t^2}{2}\right), \quad \forall t > 0, \forall i \tag{31}$$

(c) ***Restricted Isometry Property:*** *The design matrix X satisfies the RIP condition of order $2s$ with constant $\delta_{2s} \in (0, \delta^*)$, where $\delta^* < 1$ is a universal constant. Specifically, for all vectors $v \in \mathbb{R}^p$ with $\|v\|_0 \le 2s$:*

$$(1-\delta_{2s})\|v\|_2^2 \le \frac{1}{n}\|Xv\|_2^2 \le (1+\delta_{2s})\|v\|_2^2 \tag{32}$$

Theorem 3. *Under Assumption 2, consider the ℓ_1-regularized MAE estimator (least absolute deviations with Lasso penalty):*

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\|y - X\beta\|_1 + \lambda\|\beta\|_1 \right\}, \tag{33}$$

where the regularization parameter is chosen as $\lambda = C_0\sigma\sqrt{\frac{\log p}{n}}$ with $C_0 > 0$ sufficiently large.

Then, provided that n is sufficiently large and $\delta_{2s} < \delta^*$, there exist universal constants $C > 0$ and $c > 0$ such that with probability at least $1 - p^{-c}$:

$$\|\hat{\beta} - \beta^*\|_2 \le C\sigma\sqrt{\frac{s\log p}{n}} \tag{34}$$

Proof. We establish the error bound through a systematic analysis involving cone constraints, concentration inequalities, and the restricted isometry property.

Step 1: Error Decomposition and Notation.

Define the estimation error as $h = \hat{\beta} - \beta^*$. Let $S = \mathrm{supp}(\beta^*)$ denote the support set of the true parameter, with $|S| \le s$. We decompose the error vector as:

$$h = h_S + h_{S^c}, \tag{35}$$

where h_S and h_{S^c} represent the restrictions of h to the support and its complement, respectively.

Step 2: Optimality Condition Analysis.

Since $\hat{\beta}$ minimizes the objective function in (33), we have the fundamental inequality:

$$\frac{1}{n}\|y - X\hat{\beta}\|_1 + \lambda\|\hat{\beta}\|_1 \le \frac{1}{n}\|y - X\beta^*\|_1 + \lambda\|\beta^*\|_1 \tag{36}$$

Substituting the model equation $y = X\beta^* + \varepsilon$ and rearranging:

$$\frac{1}{n}\|\varepsilon - Xh\|_1 - \frac{1}{n}\|\varepsilon\|_1 + \lambda(\|\hat{\beta}\|_1 - \|\beta^*\|_1) \le 0 \tag{37}$$

Step 3: ℓ_1 Norm Relationships.

Using the triangle inequality and the decomposition $\hat{\beta} = \beta^* + h$:

$$\|\hat{\beta}\|_1 = \|\beta_S^* + h_S\|_1 + \|h_{S^c}\|_1 \tag{38}$$

$$\|\beta^*\|_1 = \|\beta_S^*\|_1 \tag{39}$$

By the reverse triangle inequality:

$$\|\beta_S^* + h_S\|_1 \ge \|\beta_S^*\|_1 - \|h_S\|_1 \tag{40}$$

Therefore:

$$\|\hat{\beta}\|_1 - \|\beta^*\|_1 \ge -\|h_S\|_1 + \|h_{S^c}\|_1 \tag{41}$$

Step 4: Concentration of the Noise Component.

Define the random vector $\nu = \frac{1}{n}X^T\varepsilon$. Under the sub-Gaussian assumption, standard concentration results yield:

$$\mathbb{P}\left(\|\nu\|_\infty \le C_1\sigma\sqrt{\frac{\log p}{n}} \right) \ge 1 - p^{-c} \tag{42}$$

for appropriate constants $C_1, c > 0$.

Step 5: Lower Bound for the Residual Term.

For any subgradient $z \in \partial\|\varepsilon\|_1$ (i.e., $\|z\|_\infty \leq 1$ and $z^T\varepsilon = \|\varepsilon\|_1$), we have:

$$\frac{1}{n}\|\varepsilon - Xh\|_1 - \frac{1}{n}\|\varepsilon\|_1 \geq \frac{1}{n}z^T(\varepsilon - Xh) - \frac{1}{n}\|\varepsilon\|_1 \quad (43)$$

$$= -\frac{1}{n}z^T(Xh) \quad (44)$$

$$= -h^T\nu_{restricted} \quad (45)$$

where $\nu_{restricted}$ represents the appropriately restricted version of ν.

Using Hölder's inequality and the concentration bound:

$$\left|h^T\nu_{restricted}\right| \leq \|h\|_1\|\nu\|_\infty \leq C_1\sigma\sqrt{\frac{\log p}{n}}\|h\|_1 \quad (46)$$

Step 6: Derivation of the Cone Constraint.

Combining the results from Steps 2-5 with inequality (37):

$$-C_1\sigma\sqrt{\frac{\log p}{n}}\|h\|_1 + \lambda(-\|h_S\|_1 + \|h_{S^c}\|_1) \leq 0 \quad (47)$$

Rearranging and using the choice $\lambda = C_0\sigma\sqrt{\frac{\log p}{n}}$ with $C_0 > 2C_1$:

$$\lambda\|h_{S^c}\|_1 \leq C_1\sigma\sqrt{\frac{\log p}{n}}\|h\|_1 + \lambda\|h_S\|_1 \quad (48)$$

$$\leq C_1\sigma\sqrt{\frac{\log p}{n}}(\|h_S\|_1 + \|h_{S^c}\|_1) + \lambda\|h_S\|_1 \quad (49)$$

$$= (C_1\sigma\sqrt{\frac{\log p}{n}} + \lambda)\|h_S\|_1 + C_1\sigma\sqrt{\frac{\log p}{n}}\|h_{S^c}\|_1 \quad (50)$$

Since $\lambda = C_0\sigma\sqrt{\frac{\log p}{n}}$ and $C_0 > 2C_1$:

$$(\lambda - C_1\sigma\sqrt{\frac{\log p}{n}})\|h_{S^c}\|_1 \leq (C_1\sigma\sqrt{\frac{\log p}{n}} + \lambda)\|h_S\|_1 \quad (51)$$

$$\frac{\lambda}{2}\|h_{S^c}\|_1 \leq 2\lambda\|h_S\|_1 \quad (52)$$

This yields the crucial cone constraint:

$$\|h_{S^c}\|_1 \leq 4\|h_S\|_1 \quad (53)$$

Step 7: Application of the Restricted Isometry Property.

The cone constraint ensures that h belongs to a restricted set where the RIP condition provides effective control. Specifically, for vectors satisfying $\|v_{S^c}\|_1 \leq 4\|v_S\|_1$, we can establish the restricted eigenvalue condition:

$$\frac{1}{n}\|Xh\|_2^2 \geq \kappa\|h\|_2^2 \quad (54)$$

where $\kappa = \frac{1-\delta_{2s}-\gamma}{2}$ for some small constant $\gamma > 0$ that depends on the cone structure.

Step 8: Final Error Bound.

From the cone constraint and Cauchy-Schwarz inequality:

$$\|h\|_1 = \|h_S\|_1 + \|h_{S^c}\|_1 \leq 5\|h_S\|_1 \quad (55)$$

$$\leq 5\sqrt{s}\|h_S\|_2 \leq 5\sqrt{s}\|h\|_2 \quad (56)$$

Using the basic inequality and concentration results:

$$\kappa\|h\|_2^2 \leq \frac{1}{n}\|Xh\|_2^2 \quad (57)$$

$$\leq 2\|\nu\|_\infty\|h\|_1 + 2\lambda\|h_S\|_1 \quad (58)$$

$$\leq 2C_1\sigma\sqrt{\frac{\log p}{n}} \cdot 5\sqrt{s}\|h\|_2 + 2\lambda\sqrt{s}\|h\|_2 \quad (59)$$

$$= (10C_1 + 2C_0)\sigma\sqrt{\frac{s\log p}{n}}\|h\|_2 \quad (60)$$

Dividing by $\|h\|_2$ and solving:

$$\|h\|_2 \leq \frac{(10C_1 + 2C_0)\sigma\sqrt{s\log p}}{\kappa\sqrt{n}} = C\sigma\sqrt{\frac{s\log p}{n}} \quad (61)$$

where $C = \frac{10C_1+2C_0}{\kappa}$ is a universal constant.

This establishes the desired estimation error bound and completes the proof. □

Remark 3. *This theorem establishes that under appropriate sparsity assumptions and design matrix conditions, the ℓ_1-regularized MAE estimator achieves minimax optimal convergence rates up to logarithmic factors. However, the analysis reveals fundamental challenges in high-dimensional settings where the ambient dimension p grows exponentially with the sample size n, highlighting the need for more sophisticated approaches in such regimes.*

4.3. Advantages of Autoregressive Models

We begin by establishing the mathematical framework for autoregressive processes. Consider a time series $\{y_t\}_{t=1}^T$ generated by an autoregressive model of order p, denoted AR(p), which satisfies the following stochastic difference equation:

$$y_t = \sum_{i=1}^{p}\beta_i y_{t-i} + \varepsilon_t, \quad t = p+1, \ldots, T, \quad (62)$$

where $\{\varepsilon_t\}_{t=1}^T$ constitutes a sequence of independent and identically distributed Gaussian random variables with $\mathbb{E}[\varepsilon_t] = 0$ and $\text{Var}(\varepsilon_t) = \sigma^2 < \infty$ for all t.

The fundamental theoretical property of autoregressive models that underlies their practical utility is encapsulated in the following theorem, which characterizes the asymptotic behavior of prediction accuracy as model complexity increases.

Theorem 4. *Let $\{y_t\}$ be generated by a stationary AR(∞) process with absolutely summable coefficients $\sum_{i=1}^{\infty} |\beta_i| < \infty$. Under standard regularity conditions for parameter identifiability and assuming sufficient sample size $T \to \infty$, the one-step-ahead prediction mean squared error of the least squares estimator $\hat{\boldsymbol{\beta}}(p) = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top$ satisfies:*

$$\lim_{p\to\infty} \mathbb{E}\left[(y_t - \hat{y}_t(p))^2\right] = \sigma^2, \tag{63}$$

where $\hat{y}_t(p)$ denotes the one-step-ahead prediction based on the AR(p) approximation.

Proof. We proceed by decomposing the prediction error into interpretable components and analyzing their asymptotic behavior.

For the AR(p) approximation, the least squares predictor is given by:

$$\hat{y}_t(p) = \sum_{i=1}^{p} \hat{\beta}_i y_{t-i}, \tag{64}$$

where $\hat{\beta}_i$ are the least squares estimates of the autoregressive coefficients.

The prediction error can be expressed as:

$$\begin{aligned} e_t(p) &= y_t - \hat{y}_t(p) \\ &= y_t - \sum_{i=1}^{p} \hat{\beta}_i y_{t-i} \\ &= \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t - \sum_{i=1}^{p} \hat{\beta}_i y_{t-i} \\ &= \sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i} + \sum_{i=p+1}^{\infty} \beta_i y_{t-i} + \varepsilon_t. \end{aligned} \tag{65}$$

This decomposition reveals three distinct sources of prediction error. The mean squared prediction error is therefore:

$$\begin{aligned} \mathbb{E}[e_t(p)^2] = \mathbb{E}&\left[\left(\sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i}\right)^2\right] \\ &+ \mathbb{E}\left[\left(\sum_{i=p+1}^{\infty} \beta_i y_{t-i}\right)^2\right] \\ &+ \mathbb{E}[\varepsilon_t^2] + 2\mathbb{E}\left[\varepsilon_t \sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i}\right] \\ &+ 2\mathbb{E}\left[\varepsilon_t \sum_{i=p+1}^{\infty} \beta_i y_{t-i}\right] \\ &+ 2\mathbb{E}\left[\sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i} \sum_{j=p+1}^{\infty} \beta_j y_{t-j}\right]. \end{aligned} \tag{66}$$

We now analyze each term systematically:

Cross-terms: Under the assumption that $\{\varepsilon_t\}$ is independent of past observations and the strong mixing conditions typical for stationary AR processes, the cross-terms involving ε_t vanish:

$$\mathbb{E}\left[\varepsilon_t \sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i}\right] = 0, \tag{67}$$

$$\mathbb{E}\left[\varepsilon_t \sum_{i=p+1}^{\infty} \beta_i y_{t-i}\right] = 0. \tag{68}$$

Estimation error term: As $T \to \infty$, the consistency of the least squares estimator under standard regularity conditions ensures:

$$\mathbb{E}\left[\left(\sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) y_{t-i}\right)^2\right] \to 0. \tag{69}$$

Approximation error term: Under the assumption of absolutely summable coefficients, as $p \to \infty$:

$$\mathbb{E}\left[\left(\sum_{i=p+1}^{\infty} \beta_i y_{t-i}\right)^2\right] \to 0. \tag{70}$$

Remaining cross-term: The mixed term between estimation and approximation errors also vanishes under appropriate conditions as both $T \to \infty$ and $p \to \infty$.

Combining these results, we obtain:

$$\lim_{p\to\infty} \mathbb{E}[e_t(p)^2] = \mathbb{E}[\varepsilon_t^2] = \sigma^2, \tag{71}$$

which completes the proof. □

This theorem establishes a fundamental result: the prediction error of an autoregressive model approaches the irreducible noise level σ^2 as the model order increases, becoming asymptotically independent of the underlying model parameters. This convergence property constitutes a key theoretical advantage of autoregressive models, demonstrating their capacity to systematically reduce prediction error through judicious increases in model complexity while maintaining statistical tractability.

4.4. Information-Theoretic Complementarity of Multiple Prediction Tasks

Building on the preceding results, we establish the complementary nature of our three prediction tasks from an information-theoretic perspective. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_K\}$ represent a sequence of visual tokens extracted from an image, where each x_i takes values in a discrete vocabulary $\mathcal{V}$. We define three distinct prediction

tasks: random token prediction, next-token prediction, and next-all token prediction.

For any given token x_i where $i \in \{1, 2, \ldots, K\}$, let $I(x_i; x_j)$ denote the mutual information between tokens x_i and x_j. We characterize the information captured by each prediction task through the following propositions.

Proposition 1 (Random Token Prediction Information)**.** *For random token prediction, the expected information gain when predicting a randomly masked token x_i given the set of unmasked tokens $x_{\mathcal{M}^c}$ is:*

$$\mathbb{E}_{i,\mathcal{M}}[I(x_i; x_{\mathcal{M}^c})] = \mathbb{E}_{i,\mathcal{M}}[H(x_i) - H(x_i|x_{\mathcal{M}^c})], \quad (72)$$

where $H(\cdot)$ denotes the Shannon entropy, $\mathcal{M} \subset \{1, 2, \ldots, K\}$ is the set of masked indices with $i \in \mathcal{M}$, and $\mathcal{M}^c$ denotes the complement of $\mathcal{M}$.

Proposition 2 (Next-Token Prediction Information)**.** *For next-token prediction, the information gain when predicting token x_i given all preceding tokens $x_{<i} = \{x_1, x_2, \ldots, x_{i-1}\}$ is:*

$$I(x_i; x_{<i}) = H(x_i) - H(x_i|x_{<i}). \quad (73)$$

Proposition 3 (Next-All Token Prediction Information)**.** *For next-all token prediction, the total information gain when predicting all future tokens $\{x_i, x_{i+1}, \ldots, x_K\}$ given tokens $x_{<i}$ is:*

$$\begin{aligned} &I(\{x_i, x_{i+1}, \ldots, x_K\}; x_{<i}) \\ &\quad = H(\{x_i, x_{i+1}, \ldots, x_K\}) \\ &\quad\quad - H(\{x_i, x_{i+1}, \ldots, x_K\}|x_{<i}). \end{aligned} \quad (74)$$

Proof of Propositions 1–3. We provide detailed proofs for each proposition.

Proof of Proposition 1: By the definition of mutual information between random variables X and Y, we have:

$$I(X; Y) = H(X) - H(X|Y).$$

For random token prediction, let i be a random variable representing the index of the masked token, and let $\mathcal{M}$ be a random variable representing the masking pattern. Then:

$$\begin{aligned} \mathbb{E}_{i,\mathcal{M}}[I(x_i; x_{\mathcal{M}^c})] &= \mathbb{E}_{i,\mathcal{M}}[H(x_i) - H(x_i|x_{\mathcal{M}^c})] \quad (75) \\ &= \mathbb{E}_{i,\mathcal{M}}[H(x_i)] - \mathbb{E}_{i,\mathcal{M}}[H(x_i|x_{\mathcal{M}^c})]. \quad (76) \end{aligned}$$

The linearity of expectation justifies the decomposition, establishing equation (72).

Proof of Proposition 2: This follows directly from the definition of mutual information. For fixed tokens x_i and $x_{<i}$:

$$I(x_i; x_{<i}) = H(x_i) - H(x_i|x_{<i}),$$

which establishes equation (73).

Proof of Proposition 3: Let $X_{\geq i} = \{x_i, x_{i+1}, \ldots, x_K\}$ denote the set of all tokens from position i onwards. By the definition of mutual information for joint random variables:

$$\begin{aligned} I(X_{\geq i}; x_{<i}) &= H(X_{\geq i}) - H(X_{\geq i}|x_{<i}) \quad (77) \\ &= H(\{x_i, x_{i+1}, \ldots, x_K\}) \quad (78) \\ &\quad - H(\{x_i, x_{i+1}, \ldots, x_K\}|x_{<i}), \quad (79) \end{aligned}$$

which establishes equation (74). □

To establish the complementarity of these prediction tasks, we analyze their information-theoretic properties:

Theorem 5 (Information Complementarity)**.** *The three prediction tasks capture distinct and complementary aspects of the visual token sequence:*

1. ***Random token prediction*** *captures non-sequential spatial correlations by maximizing*

$$\mathbb{E}_{i,\mathcal{M}}[I(x_i; x_{\mathcal{M}^c})], \quad (80)$$

which encourages bidirectional contextual understanding without dependence on token ordering.

2. ***Next-token prediction*** *captures local sequential dependencies by maximizing*

$$\sum_{i=2}^{K} I(x_i; x_{<i}), \quad (81)$$

which promotes understanding of local structural coherence following the tokenization order.

3. ***Next-all token prediction*** *captures global structure and long-range dependencies by maximizing*

$$\sum_{i=1}^{K-1} I(\{x_i, x_{i+1}, \ldots, x_K\}; x_{<i}), \quad (82)$$

which encourages comprehensive representation of hierarchical image organization.

Proof of Theorem 5. The complementarity follows from the distinct information sources each task accesses:

Disjoint Information Sources: Let $\mathcal{I}_{\text{rand}}$, $\mathcal{I}_{\text{next}}$, and $\mathcal{I}_{\text{all}}$ denote the information sets captured by random, next-token, and next-all prediction, respectively. We show these sets have minimal overlap:

1. Random token prediction accesses information $I(x_i; x_j)$ for arbitrary pairs (i, j) where $j \notin \mathcal{M}$, emphasizing non-sequential relationships.

2. Next-token prediction specifically targets $I(x_i; x_{<i})$, focusing on causal dependencies within the chosen ordering.

3. Next-all prediction captures $I(X_{\geq i}; x_{<i})$, which by the chain rule of mutual information can be decomposed as:

$$I(X_{\geq i}; x_{<i}) = I(x_i; x_{<i}) + I(x_{i+1}; x_{<i}|x_i) \quad (83)$$
$$+ \cdots + I(x_K; x_{<i}|x_i, \ldots, x_{K-1}), \quad (84)$$

revealing its emphasis on global conditional dependencies.

Complementary Coverage: The union of these information sources provides more comprehensive coverage than any individual task, as formalized in the next result. □

The total information captured by combining these tasks can be expressed as:

$$\begin{aligned}\mathcal{I}_{\text{total}} &= \alpha \cdot \mathbb{E}_{i,\mathcal{M}}[I(x_i; x_{\mathcal{M}^c})] \\ &+ \beta \cdot \sum_{i=2}^{K} I(x_i; x_{<i}) \\ &+ \gamma \cdot \sum_{i=1}^{K-1} I(\{x_i, \ldots, x_K\}; x_{<i}),\end{aligned} \quad (85)$$

where $\alpha, \beta, \gamma > 0$ are weighting parameters, and the expectations are taken over the appropriate distributions of indices and masking patterns.

Corollary 6 (Information Maximization)**.** *For appropriately chosen weights α, β, γ, maximizing the combined objective $\mathcal{I}_{total}$ yields:*

$$\mathcal{I}_{total} \geq \max\{\mathcal{I}_{rand}, \mathcal{I}_{next}, \mathcal{I}_{all}\},$$

where $\mathcal{I}_{rand}$, $\mathcal{I}_{next}$, and $\mathcal{I}_{all}$ represent the information captured by each individual task.

Proof of Corollary 6. This follows immediately from the non-negativity of mutual information and the complementary nature established in Theorem 5. Since the tasks access largely disjoint information sources, their combination provides strictly greater information coverage than any individual component. □

Remark 4. *The complementarity of these tasks ensures that maximizing the combined objective $\mathcal{I}_{total}$ enables TokenUnify to extract more comprehensive information from visual data than any single prediction task in isolation. This multi-task approach provides a more complete characterization of the underlying data distribution, leading to enhanced representation learning capabilities.*

4.5. Theoretical Analysis of Latent Manifold Structure

In this section, we provide a rigorous mathematical analysis of the latent representation space induced by the TokenUnify framework. Specifically, we demonstrate how the integration of multiple prediction objectives (random token, next-token, and next-all token prediction) shapes the geometric properties of the learned manifold, leading to a representation space that naturally accommodates both local and global aspects of neuronal morphology.

Preliminaries and Notation Let $\mathcal{X} = \{x_1, x_2, \ldots, x_K\} \in \mathbb{R}^{d \times K}$ denote a sequence of visual tokens extracted from a volumetric EM image. The model encodes these tokens into a latent space via an encoder function $f_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$, where θ represents the model parameters and $d' \ll d$ in typical applications.

We define the latent manifold $\mathcal{M}_\theta \subset \mathbb{R}^{d'}$ as the image of the encoder over all valid input tokens:

$$\mathcal{M}_\theta = \{f_\theta(x) \in \mathbb{R}^{d'} : x \in \mathcal{X}\} \quad (86)$$

This manifold is equipped with the pullback Riemannian metric g_θ induced by the Fisher information matrix of the encoder:

$$\begin{aligned}g_\theta(u, v) &= \mathbb{E}_{x \sim p_{\text{data}}}[u^T J_\theta(x)^T J_\theta(x) v] \\ &= \mathbb{E}_{x \sim p_{\text{data}}}[\langle J_\theta(x)u, J_\theta(x)v \rangle_{\mathbb{R}^{d'}}]\end{aligned} \quad (87)$$

where $J_\theta(x) = \frac{\partial f_\theta(x)}{\partial x} \in \mathbb{R}^{d' \times d}$ is the Jacobian of the encoder at input x.

Sectional Curvature Analysis Before proceeding, we establish precise definitions for the key geometric objects. Let $T_p\mathcal{M}_\theta$ denote the tangent space to $\mathcal{M}_\theta$ at point p. We define:

$$T\mathcal{M}_{\text{local}} := \text{span}\{v \in T_p\mathcal{M}_\theta : \|\nabla \mathcal{L}_{\text{random}}(p) \cdot v\| \geq \alpha\} \quad (88)$$

$$T\mathcal{M}_{\text{global}} := \text{span}\{v \in T_p\mathcal{M}_\theta : \|\nabla \mathcal{L}_{\text{next-all}}(p) \cdot v\| \geq \alpha\} \quad (89)$$

for some threshold $\alpha > 0$, where $T\mathcal{M}_{\text{random}} \subset T\mathcal{M}_{\text{local}}$ and $T\mathcal{M}_{\text{next-all}} \subset T\mathcal{M}_{\text{global}}$.

The sectional curvature $\kappa_\theta(u, v)$ of the manifold $\mathcal{M}_\theta$ for two linearly independent tangent vectors $u, v \in T_p\mathcal{M}_\theta$ is given by:

$$\kappa_\theta(u, v) = \frac{R(u, v, v, u)}{g_\theta(u, u)g_\theta(v, v) - g_\theta(u, v)^2} \quad (90)$$

where R is the Riemann curvature tensor associated with the metric g_θ.

Theorem 7 (Curvature Stratification in TokenUnify Manifolds)**.** *Under the TokenUnify framework with prediction objectives $\{\mathcal{L}_{random}, \mathcal{L}_{next}, \mathcal{L}_{next\text{-}all}\}$, the sectional curvature*

κ_θ of the learned manifold $\mathcal{M}_\theta$ exhibits systematic stratification correlated with the spatial scale of encoded features:

$$\kappa_\theta(v_1, v_2) \approx \begin{cases} O(\epsilon) & (v_1, v_2) \in T\mathcal{M}_{local} \\ -O(\delta) & (v_1, v_2) \in T\mathcal{M}_{global} \end{cases} \tag{91}$$

More precisely, there exist constants $\epsilon_{local}, \delta_{global} > 0$ such that for unit tangent vectors v_1, v_2:

$$|\kappa_\theta(v_1, v_2)| \leq \epsilon_{local} \quad \textit{when } (v_1, v_2) \in T\mathcal{M}_{random} \tag{92}$$

$$\kappa_\theta(v_1, v_2) \leq -\delta_{global} \quad \textit{when } (v_1, v_2) \in T\mathcal{M}_{next\text{-}all} \tag{93}$$

Proof. We establish this result through a three-step analysis of the manifold decomposition, local curvature computation, and global structure constraints.

Step 1: Manifold Decomposition

The TokenUnify training objective induces a natural stratification of the latent manifold. We decompose $\mathcal{M}_\theta$ into submanifolds corresponding to the dominant influence of each prediction task:

$$\mathcal{M}_\theta = \mathcal{M}_{\text{random}} \cup \mathcal{M}_{\text{next}} \cup \mathcal{M}_{\text{next-all}} \tag{94}$$

where:

$$\mathcal{M}_{\text{random}} := \{p \in \mathcal{M}_\theta : \|\nabla \mathcal{L}_{\text{random}}(p)\| > \|\nabla \mathcal{L}_k(p)\|, \forall k \neq \text{random}\} \tag{95}$$

$$\mathcal{M}_{\text{next-all}} := \{p \in \mathcal{M}_\theta : \|\nabla \mathcal{L}_{\text{next-all}}(p)\| > \|\nabla \mathcal{L}_k(p)\|, \forall k \neq \text{next-all}\} \tag{96}$$

and $\mathcal{M}_{\text{next}}$ is defined analogously.

Step 2: Local Feature Curvature Analysis

For directions v_1, v_2 associated primarily with local feature encoding (i.e., directions in $T\mathcal{M}_{\text{random}}$), the curvature tensor can be expressed as:

$$R(v_1, v_2, v_2, v_1) = \mathbb{E}_{x \sim p_{\text{data}}}[\langle \nabla_{v_1} \nabla_{v_2} f_\theta(x), \nabla_{v_2} \nabla_{v_1} f_\theta(x) \rangle] - \mathbb{E}_{x \sim p_{\text{data}}}[\langle \nabla_{[v_1, v_2]} f_\theta(x), \nabla_{[v_1, v_2]} f_\theta(x) \rangle] \tag{97}$$

The key insight is that for local feature directions, the random token prediction objective $\mathcal{L}_{\text{random}}$ encourages the encoder f_θ to behave approximately linearly within small spatial neighborhoods. This is because local patches exhibit relatively homogeneous statistical properties, leading to smooth, low-curvature encodings.

Formally, for local feature directions, we have the approximate commutativity:

$$\nabla_{v_1} \nabla_{v_2} f_\theta(x) \approx \nabla_{v_2} \nabla_{v_1} f_\theta(x) + O(\epsilon_{\text{local}}) \tag{98}$$

This implies that the Lie bracket term vanishes: $[v_1, v_2] = O(\epsilon_{\text{local}})$, and consequently:

$$R(v_1, v_2, v_2, v_1) = O(\epsilon_{\text{local}}^2) \tag{99}$$

Therefore, $\kappa_\theta(v_1, v_2) = O(\epsilon_{\text{local}})$ for directions encoding local features.

Step 3: Global Structure Curvature Analysis

For directions v_1, v_2 associated with global structure encoding (directions in $T\mathcal{M}_{\text{next-all}}$), the analysis is more intricate. The next-all token prediction objective requires the encoder to capture long-range dependencies and branching patterns in neuronal morphology.

Consider a token sequence $\{x_{\leq i}\}$ up to position i, with multiple valid continuations $\{x_{>i}^{(1)}, x_{>i}^{(2)}, \ldots, x_{>i}^{(B)}\}$ representing different possible branching structures. The encoder must satisfy two competing constraints:

Separation Constraint: Different branching patterns must be distinguishable:

$$\|f_\theta(x_{i+j}^{(a)}) - f_\theta(x_{i+j}^{(b)})\| \geq \delta > 0 \quad \forall a \neq b, \forall j > 0 \tag{100}$$

Continuity Constraint: Sequential tokens within the same branch remain close:

$$\|f_\theta(x_{i+j}^{(a)}) - f_\theta(x_{i+j-1}^{(a)})\| \leq \epsilon \quad \forall a, \forall j > 0 \tag{101}$$

These constraints necessitate a representation space with negative sectional curvature. To see this rigorously, consider the exponential map $\exp_p : T_p\mathcal{M}_\theta \to \mathcal{M}_\theta$ at a point $p = f_\theta(x_i)$ representing the branching location.

The separation constraint requires that geodesics emanating from p in different directions (corresponding to different branches) diverge at least linearly with distance. However, the continuity constraint limits the tangent space dimension available for encoding these branches.

By the Gauss-Bonnet theorem applied to geodesic triangles formed by branching paths, the requirement for exponential divergence of B branches in a d'-dimensional space with $B \gg d'$ implies:

$$\kappa_\theta(v_1, v_2) \leq -\frac{\log B}{4\pi \cdot \text{Area}(\triangle)} \leq -\delta_{\text{global}} \tag{102}$$

where $\triangle$ denotes a typical geodesic triangle in the branching region, and $\delta_{\text{global}} > 0$ depends on the branching complexity of neuronal structures.

This completes the proof of Theorem 7. □

The curvature stratification result has important implications for the representational capacity of the TokenUnify framework. The near-zero curvature in local feature directions ensures stable and efficient encoding of fine-grained morphological details, while the negative curvature in global structure directions provides the geometric flexibility necessary for representing complex branching patterns and long-range dependencies inherent in neuronal architectures.

4.6. Convergence Analysis of Alternating Optimization

We establish the convergence guarantee for our alternating optimization scheme under standard assumptions in non-convex stochastic optimization.

Theorem 8 (Formal Convergence Guarantee)**.** *Consider the alternating optimization algorithm with mode-switching probability distribution P_t at iteration t. Under the following regularity conditions:*

(A1) ***L-Smoothness:*** *There exists a constant $L > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$ and any mode $m \in \{AR, Mask\}$:*

$$\|\nabla \mathcal{L}_m(\theta) - \nabla \mathcal{L}_m(\theta')\| \leq L\|\theta - \theta'\| \tag{103}$$

(A2) ***Bounded Variance:*** *The mode-switching introduces bounded noise with variance parameter $\sigma^2_{mode} > 0$ such that:*

$$\mathbb{E}_{m_t \sim P_t}[\|\nabla \mathcal{L}_{m_t}(\theta_t) - \nabla \mathcal{L}(\theta_t)\|^2] \leq \sigma^2_{mode} \tag{104}$$

for all $t \geq 1$, where $\mathcal{L}(\theta_t) := \mathbb{E}_{m \sim P_t}[\mathcal{L}_m(\theta_t)]$.

(A3) ***Diminishing Step Size:*** *The learning rate follows the schedule $\eta_t = \eta_0(1 + \eta_0^2 L^2 t)^{-1/2}$ with $\eta_0 \in (0, 1/L]$.*

Then the sequence $\{\theta_t\}_{t=1}^T$ generated by the alternating optimization satisfies:

$$\min_{1 \leq t \leq T} \mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] \leq \frac{4(\mathcal{L}(\theta_1) - \mathcal{L}^*) + 2L\eta_0^2(1 + \sigma^2_{mode})}{\sqrt{T}} \cdot (1 + \log T + \sigma^2_{mode}) \tag{105}$$

Proof. **Step 1: Gradient Decomposition**

We introduce the natural filtration $\mathcal{F}_t = \sigma(\theta_1, \ldots, \theta_t, m_1, \ldots, m_{t-1})$ and decompose the stochastic gradient as:

$$g_t := \nabla \mathcal{L}_{m_t}(\theta_t) = \nabla \mathcal{L}(\theta_t) + \epsilon_t \tag{106}$$

where $\epsilon_t := \nabla \mathcal{L}_{m_t}(\theta_t) - \nabla \mathcal{L}(\theta_t)$ represents the mode-switching noise with $\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0$.

Step 2: One-Step Analysis

By L-smoothness and the update rule $\theta_{t+1} = \theta_t - \eta_t g_t$:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})|\mathcal{F}_t] \leq \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), -\eta_t g_t \rangle + \frac{L\eta_t^2}{2}\|g_t\|^2 \tag{107}$$

$$= \mathcal{L}(\theta_t) - \eta_t \|\nabla \mathcal{L}(\theta_t)\|^2 + \frac{L\eta_t^2}{2}\|g_t\|^2 \tag{108}$$

Expanding $\|g_t\|^2 = \|\nabla \mathcal{L}(\theta_t)\|^2 + 2\langle \nabla \mathcal{L}(\theta_t), \epsilon_t \rangle + \|\epsilon_t\|^2$ and taking conditional expectation:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})|\mathcal{F}_t] \leq \mathcal{L}(\theta_t) - \eta_t \left(1 - \frac{L\eta_t}{2}\right) \|\nabla \mathcal{L}(\theta_t)\|^2 + \frac{L\eta_t^2}{2}\mathbb{E}[\|\epsilon_t\|^2 | \mathcal{F}_t] \tag{109}$$

Step 3: Telescoping Sum

Taking total expectation and using assumption (A2), we obtain:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})] \leq \mathbb{E}[\mathcal{L}(\theta_t)] - \frac{\eta_t}{2}\mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] + \frac{L\eta_t^2 \sigma^2_{\text{mode}}}{2} \tag{110}$$

where we used the fact that $\eta_t \leq 1/L$ implies $1 - L\eta_t/2 \geq 1/2$.

Telescoping from $t = 1$ to T:

$$\sum_{t=1}^T \frac{\eta_t}{2}\mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] \leq \mathcal{L}(\theta_1) - \mathcal{L}^* + \frac{L\sigma^2_{\text{mode}}}{2}\sum_{t=1}^T \eta_t^2 \tag{111}$$

Step 4: Step Size Analysis

For the chosen step size schedule, we have the crucial bounds:

$$\sum_{t=1}^T \eta_t^2 \leq \eta_0^2 + \frac{1 + \log T}{L^2} \tag{112}$$

$$\sum_{t=1}^T \eta_t \geq \frac{\eta_0 \sqrt{T}}{\sqrt{1 + \eta_0^2 L^2 T}} \tag{113}$$

The first bound follows from the integral comparison $\sum_{t=1}^T (1 + \eta_0^2 L^2 t)^{-1} \leq 1 + \int_1^T (1 + \eta_0^2 L^2 x)^{-1} dx$, while the second uses the concavity of the square root function.

Step 5: Final Rate

Combining the telescoping bound with Jensen's inequality:

$$\min_{1 \leq t \leq T} \mathbb{E}[\|\nabla \mathcal{L}(\theta_t)\|^2] \leq \frac{2(\mathcal{L}(\theta_1) - \mathcal{L}^*) + L\eta_0^2 \sigma^2_{\text{mode}}}{\sum_{t=1}^T \eta_t} + \frac{\sigma^2_{\text{mode}}(1 + \log T)}{L \sum_{t=1}^T \eta_t} \tag{114}$$

$$\leq \frac{C(1 + \log T + \sigma^2_{\text{mode}})}{\sqrt{T}} \tag{115}$$

where the constant C depends polynomially on the problem parameters. □

4.6.1. Remarks on the Analysis

Remark 5 (Optimality)**.** *The convergence rate $O(\log T/\sqrt{T})$ is optimal for non-convex stochastic optimization, matching known lower bounds even in the single-mode case.*

Remark 6 (Mode-Switching Effect)**.** *The variance parameter σ^2_{mode} quantifies the additional difficulty introduced by alternating between training modes. When $\sigma^2_{mode} = 0$ (no mode switching), we recover the standard $O(\log T/\sqrt{T})$ rate.*

Remark 7 (Technical Innovation)**.** *Our proof technique extends classical SGD analysis to handle the mode-dependent gradient variance through careful decomposition of the noise term ϵ_t, which captures the stochastic nature of the mode selection process.*

4.7. Connecting Theory to Practice

The above theoretical results have direct implications for our model design. The limitations of MAE in high dimensions suggest that simply scaling up masked prediction models will yield diminishing returns for complex EM data. Similarly, the asymptotic optimality of autoregressive models motivates our use of Mamba-based sequence modeling, which can efficiently capture long-range dependencies. The information-theoretic complementarity of different prediction tasks justifies our unified approach that combines random, next-token, and next-all prediction objectives.

In practice, these theoretical insights translate to several key design choices in TokenUnify. We use a multi-task training objective that combines all three prediction tasks, maximizing the total information extracted from the data. We employ a Mamba-based architecture that efficiently models long-range dependencies in tokenized EM data. Additionally, we implement a progressive tokenization strategy that respects the natural structure of EM volumes.

The empirical results presented in the main paper validate these theoretical motivations, demonstrating that TokenUnify achieves superior performance and scaling properties compared to conventional approaches.

5. Social Impact and Future Work

The favorable scaling laws of TokenUnify present the opportunity to train a unified and generic visual feature extractor, which holds significant importance for visual tasks. A unified visual feature extractor can substantially reduce the cost of fine-tuning models for different visual tasks, thereby facilitating the application of visual technologies across various domains. We have currently validated the effectiveness of TokenUnify on long-sequence 3D biological images. Moving forward, we plan to further explore the performance of TokenUnify on natural images and other downstream tasks. Moreover, TokenUnify can be extended to multimodal domains such as image-text tasks [8, 21], demonstrating its utility in multimodal applications. We will also continue to investigate model lightweighting [9, 10] and efficient fine-tuning strategies [20, 24]. We believe that TokenUnify offers a promising approach for building large-scale, efficient visual pre-training models, contributing to advancements in the visual domain.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022. 4

[2] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 9:152591, 2015. 2

[3] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature Methods*, 14(2):101–102, 2017. 3, 6

[4] Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*, 2023. 4

[5] Jiuhai Chen and Jonas Mueller. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*, 2024. 4

[6] Yinda Chen, Wei Huang, Shenglong Zhou, Qi Chen, and Zhiwei Xiong. Self-supervised neuron segmentation with multi-agent reinforcement learning. In *IJCAI*, 2023. 6

[7] Yinda Chen, Wei Huang, Xiaoyu Liu, Shiyu Deng, Qi Chen, and Zhiwei Xiong. Learning multiscale consistency for self-supervised electron microscopy instance segmentation. In *ICASSP*, 2024. 6

[8] Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. *arXiv preprint arXiv:2403.15992*, 2024. 17

[9] Z Chen and L Jing. Multimodal semi-supervised learning for 3d objects. In *BMVC*, 2021. 17

[10] Zhimin Chen, Longlong Jing, Liang Yang, Yingwei Li, and Bing Li. Class-level confidence based 3d semi-supervised learning. In *WACV*, 2023. 17

[11] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1669–1680, 2018. 3, 6

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. 6

[13] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022. 6

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr

Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3, 6

[15] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *AAAI*, 2022. 6

[16] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015. 1, 3

[17] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell System Technical Journal*, 49(2):291–307, 1970. 3

[18] E. Kodak. Kodak lossless true color image suite (photocd pcd0992), 1993. Version 5. 6

[19] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017. 6

[20] Yufeng Li, Weimin Wang, Xu Yan, Min Gao, and MingXuan Xiao. Research on the application of semantic network in disease diagnosis prompts based on medical corpus. *International Journal of Innovative Research in Computer Science & Technology*, 12(2):1–9, 2024. 17

[21] Che Liu, Cheng Ouyang, Yinda Chen, Cesar César Quilodrán-Casas, Lei Ma, Jie Fu, Yike Guo, Anand Shah, Wenjia Bai, and Rossella Arcucci. T3d: Towards 3d medical image understanding through vision-language pre-training. *arXiv preprint arXiv:2312.01529*, 2023. 17

[22] Jiaxiang Liu, Jin Hao, Hangzheng Lin, Wei Pan, Jianfei Yang, Yang Feng, Gaoang Wang, Jin Li, Zuolin Jin, Zhihe Zhao, et al. Deep learning-enabled 3d multimodal fusion of cone-beam ct and intraoral mesh scans for clinically applicable tooth-bone reconstruction. *Patterns*, 4(9), 2023. 5

[23] Jiaxiang Liu, Tianxiang Hu, Yang Feng, Wanghui Ding, and Zuozhu Liu. Toothsegnet: Image degradation meets tooth segmentation in cbct images. In *ISBI*, 2023. 5

[24] Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Yang Feng, Jin Hao, Junhui Lv, and Zuozhu Liu. Parameter-efficient transfer learning for medical visual question answering. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023. 17

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 3

[26] Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, Cham, 2018. 5

[27] Philipp Schlegel, Alexander S Bates, Tejal Parag, Gregory SXE Jefferis, and Davi D Bock. Automatic detection of synaptic partners in a whole-brain drosophila em dataset. *Nature Methods*, 18(8):877–884, 2021. 1, 3

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 6

[29] Yipeng Sun, Yixing Huang, Linda-Sophie Schneider, Mareike Thies, Mingxuan Gu, Siyuan Mei, Siming Bayer, and Andreas Maier. Eagle: An edge-aware gradient localization enhanced loss for ct image reconstruction. *arXiv preprint arXiv:2403.10695*, 2024. 5

[30] Shin-ya Takemura, Yoshinori Aso, Toshihide Hige, Allan Wong, Zhiyuan Lu, C Shan Xu, Patricia K Rivlin, Harald Hess, Ting Zhao, Toufiq Parag, et al. A connectome of a learning and memory center in the adult drosophila brain. *eLife*, 6:e26975, 2017. 1, 3

[31] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *MICCAI*, 2020. 1, 3

[32] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *MICCAI*, 2024. 5