FROM LOCALIZATION TO QUANTIFICATION: STRUCTURE-GUIDED ATTRIBUTE PREDICTION UNDER WEAK SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantitative microstructural analysis often relies on instance segmentation to estimate material attributes. However, dense instance annotations are costly to obtain, whereas image-level attribute labels are widely available in the literature and technical reports. This supervision gap hinders the effective use of abundant attribute labels while preserving essential structural information. To address this challenge, we propose **S**tructure-**G**uided **A**ttribute **P**rediction (SGAP)[1], a two-stage framework that leverages scarce structural annotations alongside abundant image-level attribute labels. In the first stage, SGAP learns structural representations from a limited set of instance masks. In the second stage, a frozen structural pathway provides localization priors, while a LoRA-adapted attribute pathway extracts task-specific features for attribute prediction. SGAP uses segmentation-derived spatial priors to guide feature aggregation and predicts attributes through learned heads, without explicit per-instance geometric measurement. Experiments across two distinct scenarios and domains show that SGAP consistently outperforms both segment-then-measure approaches and direct prediction baselines, achieving state-of-the-art performance. These results support the use of segmentation-derived localization priors for attribute learning under heterogeneous supervision.

1 INTRODUCTION

Microstructural image analysis aims to quantify properties such as grain size, porosity, phase fraction, and defect content. Segmentation is often used as an intermediate step rather than being the final goal itself (Holm et al., 2020; Sosa et al., 2014). In principle, accurate segmentation should lead to accurate attribute estimates through subsequent post-processing. In practice, however, segmentation models often struggle when dense annotations are limited or when the test data differ from the training data (Holm et al., 2020; Archit et al., 2025). By contrast, sample-level attribute labels are widely available in the literature and experimental reports, whereas precise instance-level annotations remain scarce. These results support the use of segmentation-derived localization priors for attribute learning under heterogeneous supervision. Such literature-derived records provide a scalable source of material-level supervision, but generally do not specify which local image regions support each reported attribute. This imbalance motivates methods that can learn continuous attributes from abundant weak labels without relying on exhaustive segmentation annotations. More broadly, weakly supervised segmentation has demonstrated that image-level supervision can be translated into useful local spatial affinities without dense masks (Ahn & Kwak, 2018). However, when microscopy devices, imaging conditions, or microstructures change, segmentation models often fail to delineate accurate boundaries, limiting the generalizability of segment-then-quantify pipelines.

Existing methods mainly follow two directions. **Direction I** uses a segment-then-measure pipeline: instance masks are first predicted and then processed to calculate the target attributes. This approach depends heavily on accurate instance boundaries and is sensitive to missed grains, merged neighboring grains, and boundary-localization errors (Sosa et al., 2014; Tu et al., 2026). These errors often become more severe when models are transferred across materials, imaging conditions, or acquisition devices, reducing the reliability of the resulting attribute estimates. **Direction II** directly

[1]Code and datasets:`https://anonymous.4open.science/r/SGAP-D3FA1dacmlgxrzwfxyq`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

predicts attributes from microscopy images or learned microstructural representations, bypassing explicit segmentation (Choudhary et al., 2024; Murakami et al., 2026; Mann & Kalidindi, 2022). Although this approach reduces the need for dense annotations, it provides no explicit guidance on where the model should extract and aggregate attribute-relevant information. As a result, the model may rely on background textures, imaging styles, or other unstable cues that happen to correlate with the target attributes (Geirhos et al., 2020). These limitations motivate us to reconsider the role of segmentation in quantitative prediction:

Rather than requiring segmentation to produce measurement-ready instances, can it instead provide local spatial priors that guide attribute prediction to relevant regions and yield interpretable evidence paths?

To address this question, we propose **S**tructure-**G**uided **A**ttribute **P**rediction (**SGAP**), a framework that redefines the role of segmentation in quantitative prediction. Fig. 1 contrasts the two existing directions with the proposed framework. The framework first learns a structural localizer from a small number of mask annotations to identify regions relevant to the target attributes. Under the supervision of abundant attribute labels available at the image or sample level, a multi-head attribute branch extracts shared visual features and aggregates them using the foreground probabilities produced by the structural localizer as spatial weights. Following a shared-backbone, the structural pathway determines where evidence is collected, while each attribute head learns how the structurally guided representation maps to its corresponding quantitative target. To prevent weak attribute supervision from interfering with structural learning through the spatial weighting pathway, we adopt a two-stage training strategy with gradient blocking. This allows structural information to serve as a controlled spatial prior without requiring the structural pathway to produce measurement-ready instance boundaries.

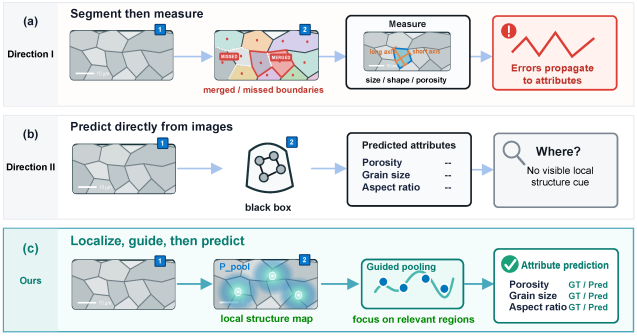


Figure 1: Comparison with existing directions.

We evaluate the proposed framework on microscopic images of ceramic and metallic materials. Compared with segment-then-measure pipelines and two vision-language models prompted to directly predict material attributes from microscopy images, structure-guided attribute prediction achieves more accurate and stable quantitative estimates. Our results support using structural information to guide feature aggregation toward regions relevant to the target attributes. In SGAP, quantitative estimates are learned from these region-guided representations under attribute supervision, rather than computed from individual instance geometries.

These results lead to four main contributions:

- We introduce a **structure-guided attribute prediction framework** for learning from scarce structural annotations and abundant weak attribute labels.
- We transform **image-segmentation features from intermediate outputs for geometric measurement into localization-level structural priors** for attribute learning.
- We systematically evaluated the effectiveness of the proposed method on datasets of different material systems and have **open-sourced our self-built ceramic Scanning Electron Microscope(SEM)-property dataset**.
- We empirically evaluate **the use of localization-level structural priors** for quantitative attribute prediction without explicit per-instance geometric measurement.

2 Related Work

2.1 Direction I: Segmentation-Based Quantification

The first direction obtains semantic or instance masks and then derives quantitative attributes from the predicted regions. General-purpose foundation models such as SAM, SAM2, and SAM3 provide transferable visual representations and promptable or concept-conditioned mask prediction (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2025). Their microscopy-oriented adaptations, including MedSAM, Segment Anything for Microscopy, and CellSAM, extend this capability to specialized imaging domains (Ma et al., 2024; Archit et al., 2025; Israel et al., 2023). For densely contacting objects, Cellpose uses a center-directed flow field to represent local instance structure, and Cellpose-SAM combines this representation with SAM-based visual encoders (Stringer et al., 2021; Pachitariu et al., 2025). These models provide the segmentation and structural representations on which downstream quantitative analysis can be built.

In the segment-then-measure pipeline, a predicted mask is treated as an intermediate geometric representation. Traditional materials-analysis systems such as MIPAR organize preprocessing, segmentation, and quantitative measurement into configurable workflows (Sosa et al., 2014). Recent studies illustrate different foundation-model-based variants: SAMM fine-tunes SAM2 for material micrographs and evaluates downstream morphology by comparing measurements computed from predicted masks with manual values (Tu et al., 2026); MatSAM uses material-aware prompt generation to guide SAM toward microstructure boundaries without additional training (Li et al., 2025), making its performance strongly dependent on prompt quality; and Mueez and Vyas explore connecting foundation-model microscopy analysis with ASTM-based grain-size estimation (Mueez & Vyas, 2026). These methods offer direct geometric basis, but the final attributes depend on the accuracy of instance boundaries, touching-object separation, and cross-domain segmentation. We retain structural prediction for localization and interpretation, while not requiring instance outputs on all literature images to reach directly measurable geometric accuracy.

2.2 Direction II: Direct Image-to-Attribute Prediction

The second direction bypasses per-instance measurement and directly learns mappings between microscopy observations or microstructure representations and material attributes. Choudhary et al. (Choudhary et al., 2024) predict grain crystallographic orientation from magnetic-domain microscopy using correlated microscopy measurements as supervision. Murakami et al. (Murakami et al., 2026) learn to predict ionic conductivity directly from SEM images of solid electrolytes. Mann and Kalidindi (Mann & Kalidindi, 2022) instead use statistical microstructure descriptors, such as two-point spatial correlations, as inputs to a CNN for modeling effective material properties. Together, these studies demonstrate that useful structure–attribute relationships can be learned without fully recovering every instance boundary.

However, direct image-to-attribute prediction with only image-level supervision does not explicitly specify which microstructural regions should provide the evidence for a prediction. We address this limitation by first learning a structural localizer with limited mask supervision and then using its spatial predictions to constrain feature aggregation during attribute learning. The resulting framework retains the scalability of direct attribute prediction while adding a spatially inspectable structural pathway. To the best of our knowledge, we are the first to introduce a structure-aware segmentation prior for image-level attribute prediction.

3 Model Architecture: Structure-Guided Two-Stage Attribute Learning

3.1 Problem Definition

We study image attribute prediction with abundant image-level labels and scarce structural annotations. Let

$$\mathcal{D}_{\mathrm{a}} = \{(\mathbf{x}_j, \mathbf{y}_j, \mathbf{v}_j, s_j)\}_{j=1}^{N_{\mathrm{a}}}, \tag{1}$$

where $\mathbf{x}_j$ is an image, $\mathbf{y}_j$ is its attribute value, $\mathbf{v}_j$ is an attribute-validity vector, with $v_{jk} = 1$ indicating that the k-th label is present; and s_j is the physical scale of the image. Missing attributes

are not filled with zeros, but are excluded from the corresponding loss through $\mathbf{v}_j$. We sample a small number of images from this dataset for manual instance annotation, yielding

$$\mathcal{D}_{\mathrm{m}} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{v}_i, s_i, \mathbf{r}_i) \mid i \in \mathcal{I}_{\mathrm{m}}\}, \qquad N_{\mathrm{m}} \ll N_{\mathrm{a}}, \tag{2}$$

where $\mathbf{r}_i$ may include instance or semantic masks. The remaining images have attribute labels but no structural annotations.

Our goal is to use the scarce structural annotations to learn a spatial prior that guides attribute prediction, while avoiding the need for measurement-grade segmentation at inference. Structural labels provide local information about object regions and boundaries, whereas image-level attributes describe global statistics without specifying their spatial origins.

We instantiate this general setting using SEM images of ceramic and metallic microstructures, with attributes such as porosity, grain size, aspect ratio, morphology, and size uniformity. Details of literature-based data construction and physical-scale calibration are provided in the Appendix.

3.2 Overall Architecture

The proposed method is a two-stage structure-guided framework for microstructure analysis and image-level attribute prediction. The model consists of a SAM3 image encoder and pixel decoder, a Cellpose-style structural head, optional semantic heads, and attribute-specific prediction heads. The structural head represents local instance organization using a foreground map and a center-pointing flow field, following the instance representation introduced in Cellpose (Stringer et al., 2021). LoRA is introduced in Stage II to adapt the visual features to image-level prediction tasks through low-rank parameter increments (Hu et al., 2022).

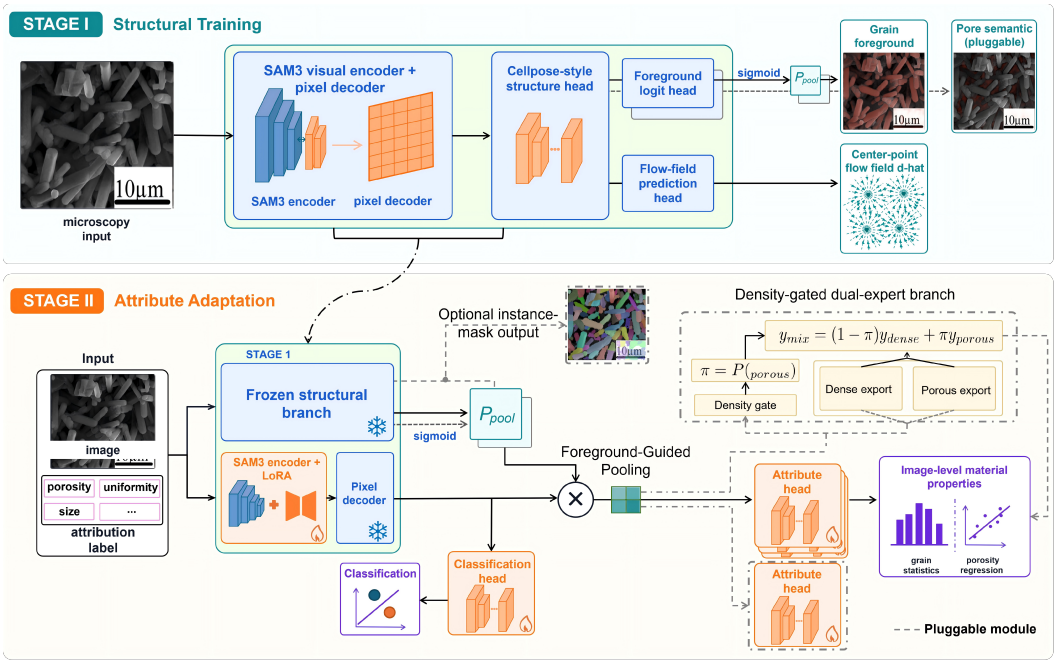


Figure 2: Overview of the proposed two-stage architecture. In Stage I, LoRA is disabled and the structural pathway is trained using pixel-level instance annotations. In Stage II, the structural pathway is frozen, while LoRA and the attribute heads are optimized using image-level supervision. Dashed components denote optional semantic or attribute-specific modules.

As shown in Fig. 2, training is divided into two stages: Stage I learns a reusable structural representation from pixel-level annotations. Stage II initializes from the Stage-I checkpoint and adapts the visual features to image-level prediction tasks. The structural pathway remains frozen and provides a stable spatial prior throughout Stage II. The dashed components indicate optional extensions: segmentation outputs can be exposed when structural inspection is required, a dual-expert branch can

be activated for attributes with distinct value regimes, and additional attribute heads can be attached when new prediction targets are introduced.

3.3 Stage I: Structural Representation Learning

Given an input image $\mathbf{x}$, the structural pathway uses the SAM3 visual encoder, feature pyramid, and pixel decoder to produce a dense spatial feature map

$$\mathbf{F}_{\mathrm{s}} = E_\theta(\mathbf{x}). \tag{3}$$

A Cellpose-style structural head then predicts a two-channel center-pointing flow field and a grain-foreground logit:

$$(\widehat{\mathbf{d}}, \widehat{\mathbf{s}}) = S_\phi(\mathbf{F}_{\mathrm{s}}). \tag{4}$$

The flow field $\widehat{\mathbf{d}}$ represents the local directions from foreground pixels toward their grain centers, while the foreground probability is obtained by applying a sigmoid function to $\widehat{\mathbf{s}}$.

During Stage I, the last two ViT blocks, the pixel decoder, and the structural head are updated using the available pixel-level instance annotations. The remaining visual-backbone parameters remain frozen. The Stage-I structural-training objective is:

$$\mathcal{L}_{\mathrm{struct}} = \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{fg}}\mathcal{L}_{\mathrm{fg}} + \mathbb{I}_{\mathrm{sem}}\lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}}, \tag{5}$$

Here, $\mathbb{I}_{\mathrm{sem}}$ indicates whether the second-phase semantic branch is enabled. Stage I also uses the available image-level attribute labels from the mask-annotated subset and the complete loss formulation is provided in the Appendix.

3.4 Stage II: Structure-Guided Attribute Prediction

Stage II starts from the Stage-I checkpoint and uses image-level attribute annotations from images without instance masks. We use LoRA to fine-tune the ViT for attribute prediction while keeping the Stage-I structural pathway fixed.

For each input, the model performs two forward passes. The LoRA-off branch preserves the feature distribution learned in Stage I and provides the structural prior, whereas the LoRA-on branch adapts the visual features to the image-level attributes:

$$\mathbf{F}_{\mathrm{off}} = E_\theta^{\text{LoRA-off}}(\mathbf{x}), \qquad \mathbf{F}_{\mathrm{on}} = E_{\theta,\Delta}^{\text{LoRA-on}}(\mathbf{x}), \qquad P_{\mathrm{pool}} = \mathrm{sg}\Big(\sigma\Big(S_\phi^{\mathrm{fg}}(\mathbf{F}_{\mathrm{off}})\Big)\Big). \tag{6}$$

Here, P_{pool} is the detached grain-foreground probability map used as the spatial weighting map for attribute pooling, S_ϕ^{fg} denotes the foreground channel of the structural head, and sg denotes stop-gradient. If enabled, the optional second-phase semantic head provides an additional region prior for pooling. These structural priors are detached from the attribute objective, allowing image-level supervision to update the attribute features without changing the structural maps that determine where evidence is collected.

The attribute heads aggregate $\mathbf{F}_{\mathrm{on}}$ using task-specific spatial weights. Grain size and aspect ratio use the grain-foreground prior, porosity uses the second-phase region prior when enabled, and morphology and size uniformity use global features. Thus, the structural pathway determines where attribute evidence is collected, while the attribute pathway learns how that evidence maps to the target attributes. Taking the present material application as an example, the porosity labels exhibit a pronounced separation between low- and high-porosity value ranges. We therefore introduce a dual-expert prediction head to model these value regimes in continuous attribute regression. This design can also be applied to other attributes with similarly multimodal distributions. Full mathematical definitions of the attribute-specific pooling weights and the density-gated dual-expert soft-mixture predictor are provided in the Appendix.

In the dual-forward pathway, the LoRA-on branch adapts the visual features to attribute prediction, while the spatial localization used for feature aggregation is provided by the LoRA-off structural pathway and inherited from the structural prior learned in Stage I. The same checkpoint can also selectively output structural segmentation results when required, allowing materials experts to inspect the predicted grain regions and overall microstructure without verifying every grain instance individually.

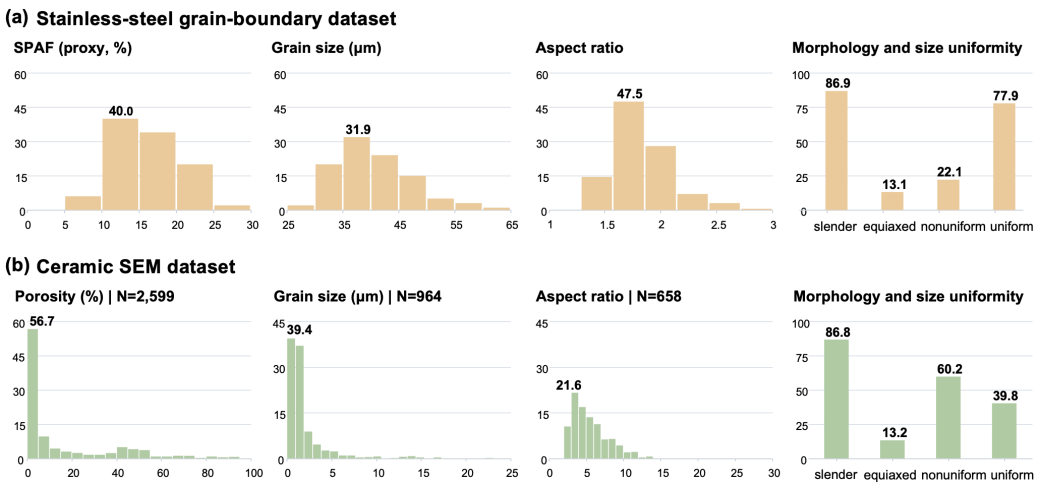


Figure 3: **Attribute distributions across the two datasets.** The orange and green rows correspond to the **Stainless-steel grain-boundary dataset** and the **Ceramic SEM dataset**, respectively. Bars show percentages among valid labels.

4 Experiments

4.1 Experimental Setup

Dataset and Tasks. We evaluate our model on two datasets: (1) a public stainless-steel grain-boundary dataset (Warren, 2023) and (2) a ceramic SEM dataset collected from the literature.

Stainless-steel grain-boundary dataset. This dataset (Warren, 2023) contains 480 images, of which 384 are used for training and 96 for testing. To better reflect real-world scenarios where densely annotated samples are scarce, we retain full expert annotations for only 100 training images, while keeping only the attribute values for the remaining 284 images.

Ceramic SEM dataset. The dataset comprises 41 expert-annotated images for Stage I, as well as 2377 training images for Stage II and 593 test images, collected from literature on silicon nitride microscopy. We adopt a publication-level split to prevent information leakage, ensuring that all images extracted from the same source paper are assigned exclusively to either the training or test set. The test set is stratified according to the availability of porosity, grain-size, and aspect-ratio labels, yielding 514, 190, and 132 valid test labels for the three attributes, respectively.

Fig. 3 summarizes the attribute distributions of these two datasets. The ceramic dataset is dominated by low-porosity samples, whereas the public dataset covers higher and more concentrated ranges of **SPAF**, grain size, and aspect ratio.

Baselines and Evaluation Protocol. We compare our method with three segment-then-measure baselines, namely SAM3, CPSAM, and MatSAM, as well as with GPT-5.5 and Gemini-3.8-Flash for direct image-to-attribute prediction. For the segmentation-based methods, the target attributes are computed from the predicted grain masks. In contrast, GPT-5.5 and Gemini-3.8-Flash directly predict the requested attributes from the microscopy images. To ensure a fair comparison, all methods use the same test-image lists, physical-scale information, and evaluation scripts for each dataset.

Implementation Details. All **SGAP** variants use the same Stage-I initialization, data splits, and random seed. Models are trained for up to 80 epochs with a batch size of 2 and four gradient-accumulation steps, with early stopping at a patience of 15 based on validation metrics. The detailed information on training and optimization is supplemented in the Appendix.

4.2 Main Results

The performance on Stainless-steel grain-boundary dataset (Warren, 2023) is as shown in Table 1, screened second-phase and defect regions are used to construct the second-phase area fraction

(**SPAF**) as an image-level proxy attribute for second-phase content. We report MAE, RMSE, MedAE, and MAPE to characterize complementary aspects of prediction error. Across the two SGAP variants, our method achieves the lowest errors among the compared methods for all SPAF and aspect-ratio metrics and for grain-size MAE, RMSE, and MAPE.

Table 1: The performance of SGAP and other methods on the public **Stainless-steel grain-boundary dataset.** The best model is in bold font, and the second-best is underlined.

Attribute	Method	MAE↓	RMSE↓	MedAE↓	MAPE(%) ↓
SPAF (proxy)	SAM3-AMG-FT	83.512	83.614	84.080	548.3
	CPSAM-FT	8.833	9.642	9.577	62.3
	MatSAM	7.153	8.242	6.513	40.1
	GPT-5.5	4.148	5.187	3.413	27.4
	Gemini-3.8-Flash	6.567	7.571	6.917	42.3
	SGAP (Ours)	**2.226**	**2.631**	**2.021**	**14.7**
	SGAP with Dual-Expert (Ours)	2.444	3.093	2.288	14.9
Grain size	SAM3-AMG-FT	37.080	38.017	37.153	90.6
	CPSAM-FT	5.416	6.847	4.294	13.4
	MatSAM	10.831	12.880	9.758	25.0
	GPT-5.5	7.998	9.708	6.538	20.0
	Gemini-3.8-Flash	5.726	7.657	**4.152**	13.9
	SGAP (Ours)	**5.186**	**6.593**	4.781	12.5
	SGAP with Dual-Expert	5.223	6.768	4.335	**12.3**
Aspect ratio	SAM3-AMG-FT	1.247	1.393	1.525	73.0
	CPSAM-FT	0.264	0.319	0.260	16.1
	MatSAM	0.421	0.477	0.388	26.1
	GPT-5.5	0.205	0.277	0.165	11.2
	Gemini-3.8-Flash	0.254	0.328	0.215	13.7
	SGAP (Ours)	0.170	0.221	**0.148**	**9.9**
	SGAP with Dual-Expert	**0.169**	**0.219**	0.157	**9.9**

We further evaluate the methods on our collected ceramic SEM dataset. As shown in Fig. 3, the prevalence of near-zero target values in this dataset causes MAPE to become highly sensitive to small denominators. We therefore supplement the ceramic evaluation with sMAPE and R^2. To better capture material-specific patterns, we enable the dual-expert branch to route model parameters according to material properties. As shown in Table 2, while the ceramic dataset yields high MAPE values across all methods, SGAP achieves substantially higher R^2 values than the baselines, most of which yielded negative R^2 values, indicates that SGAP captures meaningful attribute variation.

Table 2: The performance of SGAP and other methods on the **Ceramic SEM dataset.** The best model is in bold font, and the second-best is underlined.

Attribute	Method	MAE↓	RMSE↓	MAPE(%) ↓	sMAPE(%) ↓	R^2 ↑
Porosity	SAM3 AMG mask-FT	69.504	74.209	6161.6	148.1	-9.324
	CPSAM mask-FT	62.367	67.855	5299.2	144.6	-7.631
	MatSAM	16.022	26.717	273.0	118.8	-0.338
	GPT-5.5	15.822	24.045	668.1	111.5	-0.084
	Gemini-3.8-Flash	11.335	17.958	363.0	96.3	0.395
	SGAP (Ours)	7.809	13.963	142.7	73.2	0.635
	SGAP with Dual-Expert	**6.666**	**12.396**	**132.4**	**65.6**	**0.712**
Grain size	SAM3 AMG mask-FT	2.306	5.267	171.8	80.6	-3.360
	CPSAM mask-FT	1.405	2.628	93.9	68.5	-0.085
	MatSAM	2.314	4.366	175.0	82.9	-1.996
	GPT-5.5	1.938	3.488	150.0	74.1	-0.914
	Gemini-3.8-Flash	1.440	2.739	118.7	66.9	-0.180
	SGAP (Ours)	1.205	2.443	91.8	52.3	0.063
	SGAP with Dual-Expert	**1.078**	**2.153**	**76.0**	**51.3**	**0.272**
Aspect ratio	SAM3 AMG mask-FT	3.238	4.987	53.6	65.6	-0.392
	CPSAM mask-FT	3.226	5.015	52.9	63.4	-0.408
	MatSAM	3.079	4.947	**44.8**	60.8	-0.370
	GPT-5.5	3.091	4.591	47.7	67.1	-0.180
	Gemini-3.8-Flash	2.883	4.224	46.9	62.1	0.001
	SGAP (Ours)	**1.820**	**2.799**	46.5	**35.4**	**0.561**
	SGAP with Dual-Expert	1.934	2.895	50.2	39.1	0.531

This performance pattern is consistent with the intended role of the structural pathway as a localization-level prior rather than a measurement-ready instance segmentation module. Segment-then-measure pipelines must propagate missed, merged, or inaccurately bounded instances into subsequent geometric measurements, whereas our attribute heads predict the target properties directly from spatial features. The soft, attribute-specific pooling further concentrates feature aggregation on grain- or pore-relevant regions while retaining contextual information, allowing moderate structural errors to be less damaging than errors in a fully measured instance representation. Fig. 4 shows the

performance of SGAP with that of baseline methods on two types of datasets. Additional examples are provided in the appendix.

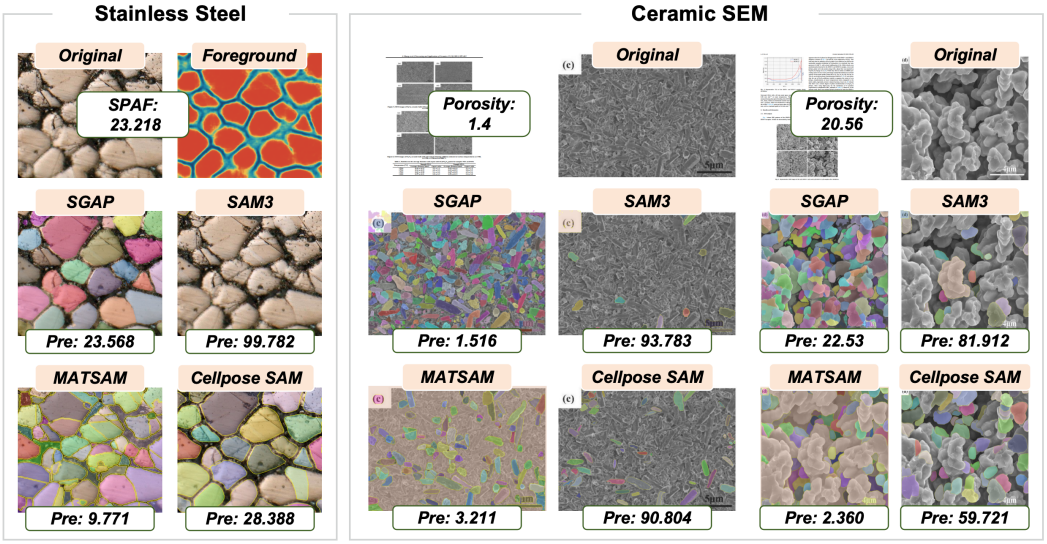


Figure 4: A case of SGAP with baseline methods on two types of datasets.

4.3 Ablation Studies

4.3.1 Benefits of Unmasked Attribute Data

To disentangle the effects of additional attribute data and visual-encoder adaptation, Fig.5 compares four training settings. *Masked-only* uses only mask-annotated images and their available attribute and classification labels, with training stopped after Stage-I. + *Unmasked Data* additionally incorporates attribute-labeled images without instance masks while keeping the ViT frozen. + *ViT FT* further fine-tunes the last two ViT blocks.

Adding unmasked attribute images improves grain-size and aspect-ratio prediction, but the gain is limited when the visual encoder is frozen. This suggests that literature-derived image-level labels complement scarce dense structural annotations, while a fixed representation cannot fully adapt to the target attribute distributions.

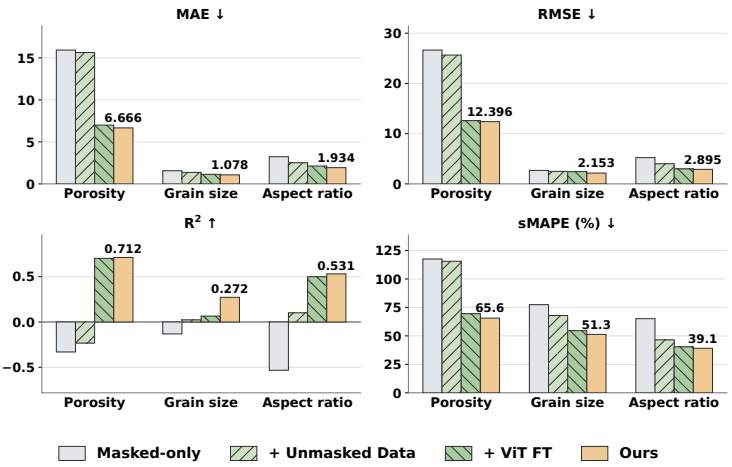


Figure 5: Effect of attribute-supervision data-use strategies.

Unfreezing selected ViT blocks yields larger gains, indicating that attribute learning requires an adaptable visual representation, but may also disturb features used for structural localization. Our method adapts the attribute pathway with LoRA while preserving a separate Stage1 structural pathway as a stable localization prior. The dual-forward and guided-pooling

design enables unmasked data to learn attribute-relevant features without compromising structural representations, demonstrating the complementarity of dense structural and image-level attribute supervision.

4.3.2 Module Ablation

Table 3 compares the complete SGAP configuration with three ablated settings. Balanced accuracy (bAcc) and macro-F1 are also reported for the imbalanced morphology and uniformity classification tasks. *w/o LoRA* disables both ViT LoRA adaptation and the separate structural pathway while retaining task-specific foreground pooling. *w/o Foreground-Guided Pathway* retains LoRA but uses a single forward pass with global average pooling instead of task-specific foreground pooling. It therefore removes both structural separation and foreground-guided aggregation, corresponding to an end-to-end image-level attribute prediction model. *Separate Structural Pathway* keeps LoRA and task-specific foreground pooling but removes the independent frozen structural pathway, so the structural prior and attribute features are obtained from the same LoRA-on forward pass. The configuration ablation examines how the attribute pathway exploits structural information. Compared with *w/o Separate Structural Pathway*, *w/o LoRA* degrades substantially on all regression and classification tasks, most notably increasing porosity MAE from 6.802 to 15.676. This result confirms that visual-representation adaptation is essential for effectively leveraging image-level attribute supervision.

Comparing *w/o Foreground-Guided Pathway* and *w/o Separate Structural Pathway* shows that task-specific foreground guidance improves grain-size prediction, particularly R^2. In contrast, global pooling is slightly more favorable for aspect-ratio prediction, suggesting that boundary-related information is important for shape attributes. Finally, the full model retains an independent frozen structural pathway as a stable localization prior while using the LoRA-adapted attribute pathway to learn task-relevant features, further improving porosity, grain-size R^2, and both classification metrics. Overall, **SGAP with full module** achieves the most balanced performance across the evaluated attributes.

Table 3: **Component-wise ablation analysis.** The full model is reported using absolute performance values. For each ablation, we report the relative performance drop with respect to the full model. For error metrics we compute $(a - f)/f$, and for higher-is-better metrics $(f - a)/f$, where a is the ablation value and f is the full-model value. Positive values (red) consistently indicate degradation, whereas negative values (blue) indicate an improvement for that individual metric.

Variant	Continuous attributes						Categorical attributes			
	Porosity		Grain size		Aspect ratio		Morphology		Uniformity	
	MAE↓	R^2 ↑	MAE↓	R^2 ↑	MAE↓	R^2 ↑	bAcc↑	F1↑	bAcc↑	F1↑
Ours	**6.666**	**0.712**	**1.078**	**0.272**	**1.934**	**0.531**	**70.68**	**0.722**	**71.34**	**0.705**
w/o LoRA	+135.2%	+130.2%	+25.0%	+86.0%	+29.5%	+80.0%	+26.9%	+28.4%	+22.8%	+26.4%
w/o Foreground-guided pathway	+2.0%	+1.1%	+1.9%	+44.9%	-3.3%	-9.4%	+3.1%	+4.7%	+14.4%	+14.6%
w/o Separate structural pathway	+2.0%	+2.0%	-1.0%	+20.6%	+15.7%	+22.0%	+2.6%	+2.5%	+7.7%	+8.2%

5 Conclusion

In this paper, we presented **SGAP**, a structure guided framework that combines structural learning from masked images with attribute supervision from unmasked literature images. Dual forward computation and attribute specific pooling transform segmentation cues into structural priors for material attribute prediction, encouraging the model to focus on grains, pores, and their spatial organization rather than irrelevant image cues.

Physical scale recovery and density gated mixing improve robustness to magnification variation and porosity prediction. Limitations include noise in literature derived labels, scale bar recognition errors, and limited mask annotations. The learned structural priors therefore provide localization and interpretive evidence rather than exact ground truth for three dimensional attributes or complete instance geometry.

6 Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets have been made publicly available in an anonymous repository[1] to facilitate replication and verification. We believe these measures will enable other researchers to reproduce our work and further advance the field.

7 AI Use Statement

During the preparation of this manuscript and its accompanying code, large language models (LLMs) were used in a limited capacity to assist with linguistic refinement, grammar correction, and improving the clarity and cohesion of the English prose. LLMs were not used to formulate the research questions, develop the scientific methodology, design the experiments, generate scientific claims or fabricate data. All factual statements, numerical results and interpretations have been manually reviewed and verified by the authors. This use of LLMs complies with prevailing academic publishing policies concerning AI-assisted writing, and the authors take full responsibility for the final content of this work.

References

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Ahn_Learning_Pixel-Level_Semantic_CVPR_2018_paper.html.

Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, Genevieve Buckley, Sebastian von Haaren, Sagnik Gupta, Marian Grade, Matthias Wirth, Gunter Schneider, Andreas Dengel, Sheraz Ahmed, and Constantin Pape. Segment anything for microscopy. *Nature Methods*, 22(3):579–591, 2025. doi: 10.1038/s41592-024-02580-4. URL https://www.nature.com/articles/s41592-024-02580-4.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Radle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719. OpenReview record for ICLR 2026.

Amit Kumar Choudhary, Tvrtko Grubesa, Andreas Jansche, Timo Bernthaler, Dagmar Goll, and Gerhard Schneider. Deep learning and correlative microscopy for quantification of grain orientation in sintered FeNdB-type permanent magnets by domain pattern analysis. *Acta Materialia*, 264: 119563, 2024. doi: 10.1016/j.actamat.2023.119563. URL https://doi.org/10.1016/j.actamat.2023.119563.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Elizabeth A. Holm, Ryan Cohn, Nan Gao, Andrew R. Kitahara, Thomas P. Matson, Bo Lei, and Srujana Rao Yarasi. Overview: Computer vision and machine learning for microstructural characterization and analysis. *Metallurgical and Materials Transactions A*, 51(12):5985–5999, 2020. doi: 10.1007/s11661-020-06008-4. URL https://link.springer.com/article/10.1007/s11661-020-06008-4.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Uriah Israel, Markus Marks, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, Caitlin Brown, Edward Pao, Shenyi Li, Alexander Pearson-Goulart, Pietro Perona, Georgia Gkioxari, Ross Barnowski, Yisong Yue, and David Van Valen. CellSAM: A foundation model for cell segmentation. *bioRxiv*, 2023. doi: 10.1101/2023.11.17.567630. URL https://www.biorxiv.org/content/10.1101/2023.11.17.567630. Preprint.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Changtai Li, Xu Han, Chao Yao, Yu Guo, Zixin Li, Lei Jiang, Wei Liu, Haiyou Huang, Huadong Fu, and Xiaojuan Ban. A novel training-free approach to efficiently extracting material microstructures via visual large model. *Acta Materialia*, 290:120962, 2025. doi: 10.1016/j.actamat.2025.120962. URL https://doi.org/10.1016/j.actamat.2025.120962.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. doi: 10.1038/s41467-024-44824-z. URL https://doi.org/10.1038/s41467-024-44824-z.

Andrew Mann and Surya R. Kalidindi. Development of a robust CNN model for capturing microstructure–property linkages and building property closures supporting material design. *Frontiers in Materials*, 9:851085, 2022. doi: 10.3389/fmats.2022.851085. URL https://doi.org/10.3389/fmats.2022.851085.

Abdul Mueez and Shruti Vyas. Bridging foundation models and ASTM metallurgical standards for automated grain size estimation from microscopy images. *arXiv preprint arXiv:2604.18957*, 2026. URL https://arxiv.org/abs/2604.18957. Preprint.

Kento Murakami, Yudai Yamaguchi, Yo Kato, Kazuki Ishikawa, Naoto Tanibata, Hayami Takeda, Masanobu Nakayama, and Masayuki Karasuyama. Deep learning based SEM image analysis for predicting ionic conductivity in $LiZr_2(PO_4)_3$-based solid electrolytes. *Digital Discovery*, 5(1):453–462, 2026. doi: 10.1039/D5DD00232J. URL https://doi.org/10.1039/D5DD00232J.

Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, 2025. doi: 10.1101/2025.04.28.651001. URL https://www.biorxiv.org/content/10.1101/2025.04.28.651001v1. Preprint.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Radle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

John M. Sosa, Daniel E. Huber, Brian Welk, and Hamish L. Fraser. Development and application of MIPAR: A novel software package for two- and three-dimensional microstructural characterization. *Integrating Materials and Manufacturing Innovation*, 3(1):123–140, 2014. doi: 10.1186/2193-9772-3-10. URL https://link.springer.com/article/10.1186/2193-9772-3-10.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021. doi: 10.1038/s41592-020-01018-x. URL https://www.nature.com/articles/s41592-020-01018-x.

Jiahao Tu, Zi Wang, Weifu Li, Liming Tan, Lan Huang, and Feng Liu. SAMM: A general-purpose segmentation model for material micrographs based on the segment anything model 2. *Advanced Powder Materials*, 5:100404, 2026. doi: 10.1016/j.apmate.2026.100404. URL https://doi.org/10.1016/j.apmate.2026.100404.

Peter Warren. Artificial grains and real grains, 2023. URL https://www.kaggle.com/datasets/peterwarren/voronoi-artificial-grains-gen.

A Additional Details and Results

A.1 Additional Visual Comparison Examples

SGAP can optionally display structural probability maps, flow fields, and post-processed segmentation results to facilitate visual inspection by experts. Instance-mask post-processing is performed only when structural output is requested. As shown in Fig. 6, we present further visual comparisons of segmentation results, covering the two types of datasets used in the main experiments. Each case includes the original image, the predicted foreground probability, our structural output, and corresponding outputs from SAM3, MatSAM, and Cellpose-SAM. The examples cover different microstructural appearances and object-contact conditions. Compared with dedicated baseline segmentation models, SGAP achieved superior segmentation performance and more accurate attribute prediction by leveraging weak supervision from attribute data, even with limited annotated segmentation data. These results illustrate the inspectability of the optional structural pathway and provide qualitative evidence for the benefits of structure-guided attribute prediction.

A.2 Literature Retrieval and Expert-Validated Dataset Construction

We constructed the ceramic dataset using 1639 documents on silicon nitride materials retrieved via relevant keywords provided by domain experts. We collect the relevant passages, microscopy figures, figure captions, and tables from these sources. The attribute extraction pipeline initially yields 5884 image-level records. After physical-scale calibration and label-quality filtering, 3011 images and their associated data are retained. Domain experts select 41 of these images for manual structural annotation. Therefore, the final dataset comprises 2970 records linking images to their attributes, along with 41 structural images and their corresponding attribute data, all annotated by experts.

To ensure that the dataset is accepted on the basis of automated extraction, domain experts participate in the construction and validation process. For each candidate image, a prompt-based large language model extracts the available material attributes from the collected textual and visual context. Two associate professors and a doctoral students with relevant domain knowledge jointly establish and refine the extraction prompt, including the definitions of the target fields, admissible value ranges, and criteria for distinguishing mean grain size, grain aspect ratio, porosity, morphology, and size uniformity from unrelated quantities. A manual audit of sampled extraction records yields an accuracy of 89.1% under the adopted audit criteria, providing a direct estimate of the reliability of the extracted attributes. Missing attributes remain explicitly marked as unavailable rather than being imputed as numerical zeros, and an attribute contributes to training only when its validity indicator is active.

For structural validation, domain experts select and manually annotate 41 microscopy images. They delineate the target grain instances and, where applicable, the second-phase or pore regions, and use Fiji to calibrate the physical scale from the scale bar and to inspect and measure the annotated image data. These expert-calibrated annotations form a high-confidence structural subset for learning local responses and spatial regions. The remaining literature images retain their image-level attributes but do not require instance masks, allowing the attribute stage to use a substantially larger set of records. Together, expert-defined extraction criteria, sampled manual auditing, and physically calibrated structural annotations provide complementary evidence for the scientific reliability of the collected dataset.

The visual masks generated from the structural subsets annotated by experts are shown in the figure. These examples display manually delineated grain instances, as well as regions representing

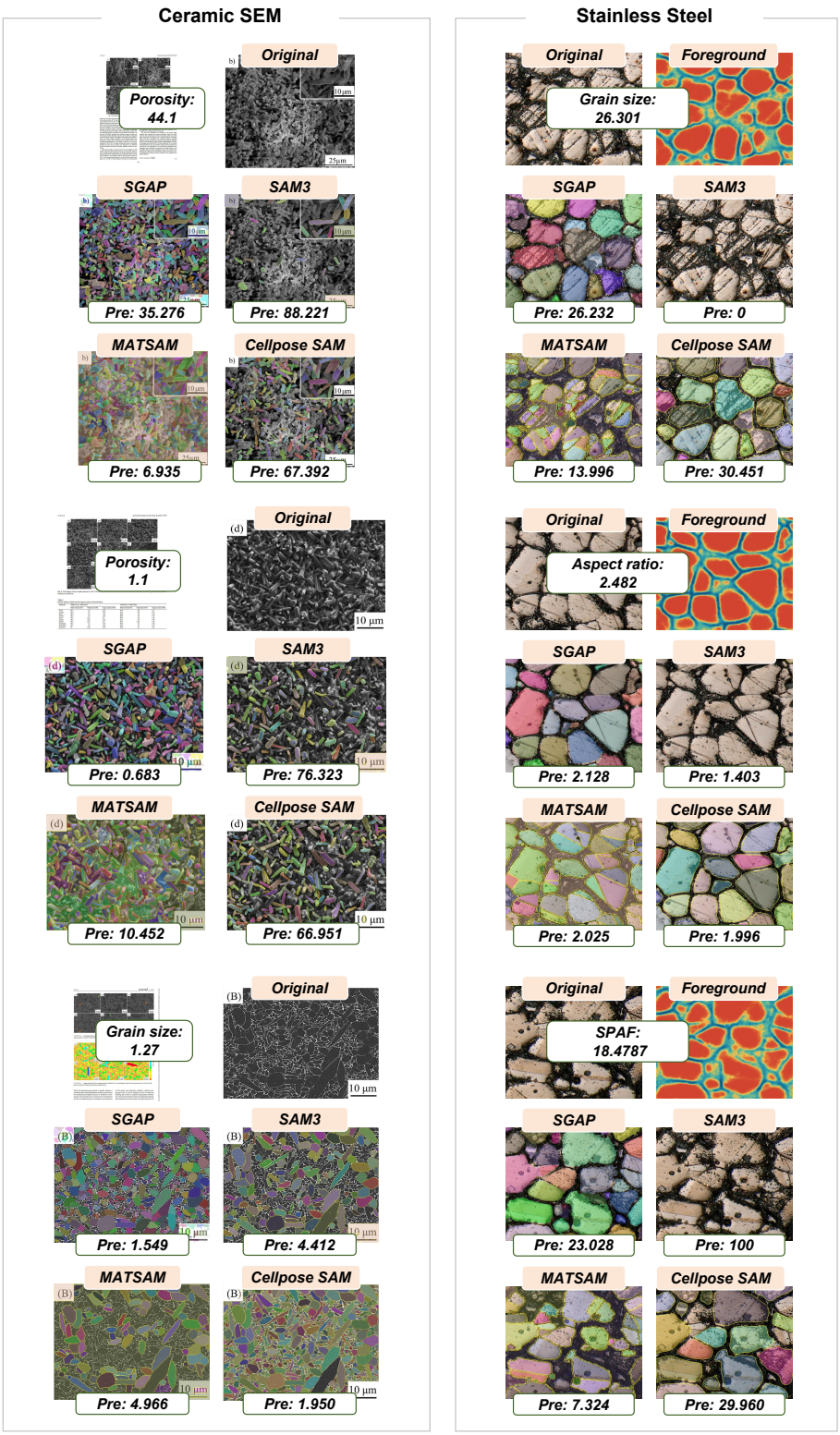
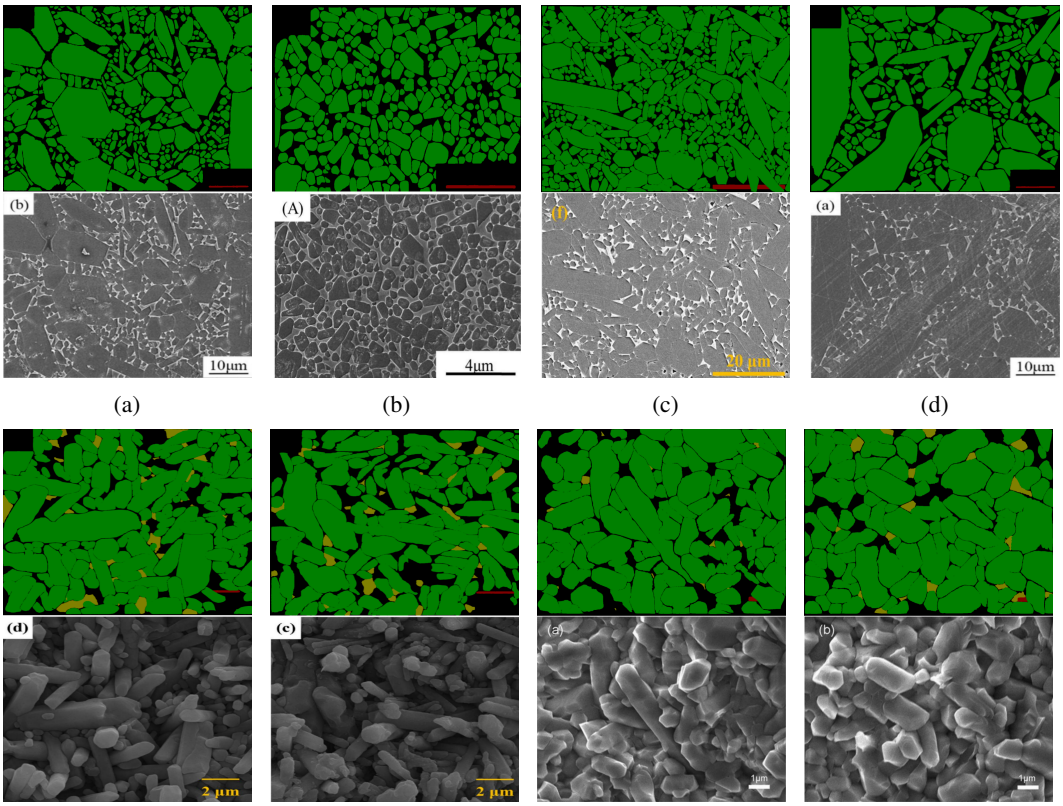


Figure 6: Additional examples of visual comparisons with the baseline across two datasets.

secondary phases or pores (depending on the specific case) used for structural supervision and validation.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: Visualization of mask annotations derived from representative microscopy images provided by experts; above is the mask annotation; below is the original image.

A.3 Dataset Construction and Attribute Records

Each record stores the available image-level attributes and physical scale. The masked subset additionally provides grain-instance annotations for Stage I, while the unmasked subset contributes only attribute and classification supervision.

For the public stainless-steel grain-boundary dataset, direct second-phase labels are unavailable. We therefore use the screened second-phase and defect regions from the original micrographs to define an image-level second-phase area fraction (SPAF) proxy:

$$\mathrm{SPAF}(\mathbf{x}) = 100 \times \frac{|\mathcal{R}_{\mathrm{sp}}(\mathbf{x})|}{|\Omega(\mathbf{x})|}, \tag{7}$$

where $\mathcal{R}_{\mathrm{sp}}$ denotes the screened second-phase or defect region and Ω is the evaluated image region. SPAF is used only as an image-level proxy target and is not interpreted as the true three-dimensional porosity of the material.

Ceramic porosity MAPE is high because the reported bulk measurement need not correspond pixel-wise to pores in a local micrograph, and MAPE amplifies errors near zero. Samples with porosity below 5% account for approximately 56.7% of the valid records. We therefore treat MAPE as supplementary and use sMAPE and R^2 as the primary ceramic metrics.

A.4 SAM-Style Visual Foundation Models

Segment Anything (SAM) introduced a general-purpose visual foundation model for promptable segmentation. It was pretrained on a large and diverse collection of images and masks, enabling a common visual representation to be transferred to new segmentation tasks with lightweight prompting or adaptation (Kirillov et al., 2023). SAM2 extended this framework from static images to images and videos by incorporating memory-based mechanisms that maintain object information

across frames (Ravi et al., 2025). SAM3 further generalizes the foundation-model interface to concept-driven object detection, segmentation, and tracking, allowing visual concepts to serve as the basis for object-level predictions (Carion et al., 2025).

Several subsequent studies adapted this family of models to specialized domains. MedSAM adapts SAM to medical image segmentation, while Segment Anything for Microscopy and CellSAM investigate domain-specific adaptation for microscopy and cellular images, respectively (Ma et al., 2024; Archit et al., 2025; Israel et al., 2023). These methods show that the transferable visual representations learned by general segmentation foundation models can provide useful initialization for images with domain-specific appearance, object scales, and boundary characteristics.

In this work, SAM3 is used primarily as a dense visual representation rather than as a standalone promptable segmentation system. The SAM3 image encoder and pixel decoder produce a spatial feature map, from which a Cellpose-style structural head predicts a grain-foreground logit and a two-dimensional center-pointing flow field. This design uses the local structural representation of Cellpose (Stringer et al., 2021) while retaining the transferable visual features of a segmentation foundation model. The structural head is trained with a small number of grain-instance masks in Stage I. Its foreground response is subsequently used as a detached spatial prior for attribute-specific feature pooling in Stage II.

The implementation does not require text prompts or concept-conditioned fusion during the forward pass. Instead, the foundation-model feature map is shared by the structural and attribute pathways: the LoRA-off path preserves the Stage-I structural response, whereas the LoRA-on path adapts visual features to image-level material attributes. Thus, the role of the SAM3 backbone in our framework is to provide transferable dense visual features and a spatial basis for structure-guided attribute prediction, rather than to produce measurement-ready instance masks for every image.

A.5 Structural Targets and Stage-I Losses

Let $\mathbf{m}_i$ denote the binary mask of instance i. The foreground target is the union of all instances,

$$m_u^* = \mathbb{I}\left(\sum_i m_{i,u} > 0\right). \tag{8}$$

For each instance mask $\mathbf{m}_i$, we construct a scalar potential h_i by iteratively diffusing mass from an interior reference point. The target flow at pixel u is defined as the normalized spatial gradient of this potential:

$$\mathbf{d}_{i,u}^* = \frac{\nabla_u h_i(u)}{\|\nabla_u h_i(u)\|_2 + \epsilon}. \tag{9}$$

The instance-wise fields are assembled into a two-channel flow target $\mathbf{d}^*$, with $\mathbf{d}_u^* = \mathbf{0}$ at background pixels. The resulting vectors provide local directions toward instance centers without requiring a discrete instance identifier at every pixel.

The structural loss combines a full-image flow MSE, where the normalized flow target is scaled by a factor of five and the mean two-channel error is divided by two, with binary cross-entropy applied directly to the foreground logits:

$$\mathcal{L}_{\text{flow}} = \frac{1}{4|\Omega|}\sum_{u\in\Omega}\left\|\widehat{\mathbf{d}}_u - 5\mathbf{d}_u^*\right\|_2^2, \qquad \mathcal{L}_{\text{fg}} = \frac{1}{|\Omega|}\sum_{u\in\Omega}\text{BCEWithLogits}(\widehat{s}_u, m_u^*). \tag{10}$$

The flow loss is computed over all pixels, with the background flow target equal to zero, while the foreground loss is also computed over all pixels. If second-phase masks are available, the optional semantic loss is a masked binary cross-entropy:

$$\mathcal{L}_{\text{sem}} = \frac{1}{\sum_i a_i + \epsilon}\sum_i a_i \frac{1}{|\Omega_i|}\sum_{u\in\Omega_i}\text{BCEWithLogits}(\widehat{q}_{i,u}, q_{i,u}^*), \tag{11}$$

where $a_i = 1$ if the second-phase mask for image i is present and 0 otherwise. The semantic branch is therefore optional and can be extended to other binary region labels without changing the grain flow representation. We first collect the structure-specific terms as

$$\mathcal{L}_{\text{struct}} = \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{fg}}\mathcal{L}_{\text{fg}} + \mathbb{I}_{\text{sem}}\lambda_{\text{sem}}\mathcal{L}_{\text{sem}}, \tag{12}$$

The Stage-I training records retain their available image-level attribute and classification labels. The optional density-gated porosity branch contributes only its own gate and regression terms. We define

$$\mathcal{L}_{\mathrm{por}} = \lambda_{\mathrm{gate}} \langle \mathrm{CE}(\mathbf{g}_b, c_b) \rangle_{\mathcal{V}} + \lambda_{\mathrm{attr}} \lambda_{\mathrm{exp}} \sum_{r \in \{\mathrm{l},\mathrm{h}\}} \langle \ell_{\mathrm{SL1}}(\widehat{z}_b^{\,r}, z_b^r) \rangle_{\mathcal{V}_r} + \lambda_{\mathrm{attr}} \lambda_{\mathrm{mix}} \left\langle \ell_{\mathrm{SL1}}(\widehat{z}_b^{\,\mathrm{mix}}, z_b) \right\rangle_{\mathcal{V}}. \tag{13}$$

Here $\mathcal{V}$ contains valid porosity or SPAF labels and $\mathcal{V}_r$ contains valid labels in regime $r \in \{\mathrm{l}, \mathrm{h}\}$, as defined in the dual-expert subsection below; ℓ_{SL1} is the elementwise Smooth L1 loss. Thus, $\mathcal{L}_{\mathrm{por}}$ does not include $\mathcal{L}_{\mathrm{attr}}^{(-\mathrm{sp})}$ or $\mathcal{L}_{\mathrm{cls}}$. The complete Stage-I objective is therefore

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{struct}} + \mathcal{L}_{\mathrm{attr}}^{(\mathrm{I})} + \mathcal{L}_{\mathrm{cls}} + \mathbb{I}_{\mathrm{por}} \mathcal{L}_{\mathrm{por}}, \tag{14}$$

where $\mathbb{I}_{\mathrm{por}}$ indicates whether the optional branch is enabled, $\mathcal{L}_{\mathrm{attr}}^{(\mathrm{I})} = \mathcal{L}_{\mathrm{attr}}$ when it is disabled, and $\mathcal{L}_{\mathrm{attr}}^{(\mathrm{I})} = \mathcal{L}_{\mathrm{attr}}^{(-\mathrm{sp})}$ otherwise. The attribute, classification, and porosity-branch terms are defined in Eqs. 20, 21, and 13, respectively; unavailable image-level labels are excluded by their validity masks. During Stage I, the final two ViT blocks, pixel decoder, structural head, and any enabled semantic head are trainable. The remaining SAM3 parameters are frozen, including the feature-pyramid parameters. This partial unfreezing allows the dense feature map to adapt to the microscopy domain while retaining most of the pretrained visual backbone.

A.6 Stage-II Feature Adaptation and Dual Forward

Stage II starts from the Stage-I checkpoint. A LoRA module wraps selected ViT linear layers. For a frozen linear mapping $\mathbf{W}$, the adapted mapping is

$$\mathbf{W}'\mathbf{h} = \mathbf{W}\mathbf{h} + \frac{\alpha}{r}\mathbf{B}\mathbf{A}\mathbf{h}, \tag{15}$$

where r is the low-rank dimension, α is the scaling parameter, and $\mathbf{A}$ and $\mathbf{B}$ are trainable. The base structural pathway remains frozen while the LoRA parameters and attribute heads are updated. The dual-forward computation is defined in Eq. 6; if the optional second-phase semantic head is enabled, its detached prior is

$$P_{\mathrm{sp}} = \mathrm{sg}\Big(\sigma\Big(S_\phi^{\mathrm{sp}}(\mathbf{F}_{\mathrm{off}})\Big)\Big). \tag{16}$$

Both structural maps are resized to the spatial resolution of $\mathbf{F}_{\mathrm{on}}$ before pooling. When the model is run without the dual-forward option, a single forward pass is used; the attribute and structural heads then read the same dense feature map.

A.7 Attribute-Specific Pooling and Prediction Heads

We construct global, grain-guided, and second-phase-region-guided representations for attributes associated with different spatial evidence.

The grain-foreground prior $P^{\mathrm{g}} \equiv P_{\mathrm{pool}}$ and optional second-phase prior P^{sp} are produced by the LoRA-off structural pathway and resized to the feature grid. For feature vector $\mathbf{F}_{\mathrm{on},i}$ at location i, the representations are

$$\begin{aligned} \overline{\mathbf{f}}_{\mathrm{G}} &= \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbf{F}_{\mathrm{on},i}, \\ \overline{\mathbf{f}}_{\mathrm{grain}} &= \frac{\sum_{i \in \Omega} P_i^{\mathrm{g}} \mathbf{F}_{\mathrm{on},i}}{\sum_{i \in \Omega} P_i^{\mathrm{g}} + \epsilon}, \\ \overline{\mathbf{f}}_{\mathrm{sp}} &= \frac{\sum_{i \in \Omega} P_i^{\mathrm{sp}} \mathbf{F}_{\mathrm{on},i}}{\sum_{i \in \Omega} P_i^{\mathrm{sp}} + \epsilon}. \end{aligned} \tag{17}$$

Here, Ω contains all feature-grid locations; the guided representations are continuous soft averages and require no thresholding or instance clustering.

Each attribute selects the representation associated with its spatial evidence:

$$\mathbf{f}_k = \begin{cases} \overline{\mathbf{f}}_{\mathrm{grain}}, & k \in \{\text{grain size}, \text{aspect ratio}\}, \\ \overline{\mathbf{f}}_{\mathrm{sp}}, & k \in \{\text{porosity}, \text{density gate}, \text{dense expert}, \text{porous expert}\}, \\ \overline{\mathbf{f}}_{\mathrm{G}}, & k \in \{\text{morphology}, \text{size uniformity}\}. \end{cases} \tag{18}$$

If the second-phase prior is unavailable, the corresponding attribute head falls back to global pooling. The second-phase map is used only as a spatial prior; the attribute head still regresses the reported image-level target.

A.8 Missing Labels and Attribute Losses

For continuous attribute k, each record stores a validity indicator v_{jk}. A missing label is represented internally by a placeholder value but has $v_{jk} = 0$ and does not contribute to the loss. After any enabled target transformation, the valid values are min-max normalized using the range used by the corresponding training run:

$$\widetilde{y}_{jk} = \frac{y_{jk} - a_k}{b_k - a_k}, \qquad a_k = \min_{j:v_{jk}=1} y_{jk}, \quad b_k = \max_{j:v_{jk}=1} y_{jk}. \tag{19}$$

The masked Smooth L1 objective is

$$\mathcal{L}_{\text{attr}} = \lambda_{\text{attr}} \frac{\sum_{j,k} v_{jk}\, \text{SmoothL1}(\widehat{\widetilde{y}}_{jk}, \widetilde{y}_{jk})}{\sum_{j,k} v_{jk} + \epsilon}. \tag{20}$$

Categorical attributes use an analogous per-sample validity mask. For class c, with logits $\widehat{\mathbf{q}}_{jc}$ and validity r_{jc},

$$\mathcal{L}_{\text{cls}} = \lambda_{\text{cls}} \sum_c \frac{\sum_j r_{jc}\, \text{CE}(\widehat{\mathbf{q}}_{jc}, q_{jc})}{\sum_j r_{jc} + \epsilon}, \tag{21}$$

with inverse-frequency class weights computed on the training split.

Any target transformation is inverted before evaluation, so all reported metrics remain in the original physical units.

A.9 Optional Regime-Gated Dual-Expert Regression

The same optional dual-expert module is applied to the second-phase-related percentage target in both datasets. We denote this target by y_b^{sp}: it is the reported porosity for the collected ceramic dataset and the image-derived SPAF proxy for the public dataset. These labels have different physical meanings, but they use the same prediction architecture and loss. The module takes the second-phase-region-guided representation $\bar{\mathbf{f}}_{\text{sp}}$ as input and uses three independent two-layer MLPs to produce gate logits $\mathbf{g}_b$, a low-regime expert output $\widehat{z}_b^{\text{l}}$, and a high-regime expert output $\widehat{z}_b^{\text{h}}$ for sample b. Each MLP has a 256-dimensional hidden layer with ReLU and dropout of 0.1. Given the shared threshold $\tau = 10\%$, the ground-truth regime is $c_b = \mathbb{I}[y_b^{\text{sp}} \geq \tau]$, and the predicted probability of the high-value regime is

$$\pi_b = \text{softmax}(\mathbf{g}_b)_{\text{high}}. \tag{22}$$

The low- and high-regime experts are trained against targets normalized within their respective value ranges. After subrange denormalization and the inverse of the dataset-specific target transformation T_{sp}, when one is enabled, the two outputs give percentage predictions $\widehat{y}_b^{\text{l}}$ and $\widehat{y}_b^{\text{h}}$. The prediction used at inference is their probability-weighted mixture:

$$\widehat{y}_b^{\text{mix}} = (1 - \pi_b)\widehat{y}_b^{\text{l}} + \pi_b \widehat{y}_b^{\text{h}}. \tag{23}$$

Mixing in percentage space gives the interpolation a direct interpretation in the reported units. The result is subsequently transformed and globally normalized using the same mapping as the training target.

Let z_b^r be the within-regime normalized target for expert $r \in \{\text{l}, \text{h}\}$, let z_b be the globally normalized second-phase-related target, and let $\widehat{z}_b^{\text{mix}} = \mathcal{N}(T_{\text{sp}}(\widehat{y}_b^{\text{mix}}))$. We write $\langle\cdot\rangle_{\mathcal{A}}$ for an average over a set of valid samples, with $\mathcal{V}$ denoting all samples with a valid porosity or SPAF label, $\mathcal{V}_{\text{l}} = \{b \in \mathcal{V} : c_b = 0\}$, and $\mathcal{V}_{\text{h}} = \{b \in \mathcal{V} : c_b = 1\}$. The implemented Stage-II objective is

$$\begin{aligned} \mathcal{L}_{\text{II}} = \mathcal{L}_{\text{attr}}^{(-\text{sp})} + \mathcal{L}_{\text{cls}} + \lambda_{\text{gate}} \langle \text{CE}(\mathbf{g}_b, c_b) \rangle_{\mathcal{V}} \\ + \lambda_{\text{attr}} \lambda_{\text{exp}} \sum_{r \in \{\text{l},\text{h}\}} \langle \ell_{\text{SL1}}(\widehat{z}_b^r, z_b^r) \rangle_{\mathcal{V}_r} \\ + \lambda_{\text{attr}} \lambda_{\text{mix}} \left\langle \ell_{\text{SL1}}(\widehat{z}_b^{\text{mix}}, z_b) \right\rangle_{\mathcal{V}}, \end{aligned} \tag{24}$$

Here, the gate is supervised by the low/high regime class, each expert is directly supervised only by samples from its corresponding regime, and the mixture loss matches the soft route used at inference. The latter also allows the final regression error to update the gate and, when their routing probabilities are nonzero, both experts, thereby reducing discontinuities caused by hard routing near the threshold. The term $\mathcal{L}_{\text{attr}}^{(-\text{sp})}$ denotes Eq. 20 with the dataset's second-phase-related target masked out, preventing the original single regressor from receiving a competing loss for porosity or SPAF. The weights are $(\lambda_{\text{attr}}, \lambda_{\text{cls}}, \lambda_{\text{gate}}, \lambda_{\text{exp}}, \lambda_{\text{mix}}) = (5.0, 0.1, 0.5, 1.0, 0.3)$. `SGAP (Ours)` disables this optional branch and uses $\mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{cls}}$, whereas `SGAP with Dual-Expert (Ours)` uses the complete objective above.

A.10 Physical-Scale Calibration and Unit Recovery

For each image, the physical scale is obtained from the scale-bar value and its measured pixel length:

$$s_j^{\text{orig}} = \frac{\ell_j^{\text{phys}}}{\ell_j^{\text{px}}} \quad (\mu\text{m/pixel}). \tag{25}$$

Let (W_j, H_j) be the original image dimensions. For aspect-preserving letterbox resizing to side length $L = 1008$, the resizing factor and the physical scale on the input grid are

$$\gamma_j = \frac{L}{\max(W_j, H_j)}, \qquad s_j^{\text{lb}} = \frac{s_j^{\text{orig}}}{\gamma_j}. \tag{26}$$

During training, the reported mean grain size $y_j^{\mu\text{m}}$ is converted to its pixel-space target on the input grid; after inference, the predicted value is converted back to micrometers:

$$y_j^{\text{px}} = \frac{y_j^{\mu\text{m}}}{s_j^{\text{lb}}}, \qquad \widehat{y}_j^{\mu\text{m}} = \widehat{y}_j^{\text{px}} s_j^{\text{lb}}. \tag{27}$$

The network therefore learns an apparent pixel size consistent with image resizing, while the per-image scale provides a deterministic conversion between pixel and physical units.

A.11 Prompts

Data Extraction Prompt. The following prompt was used for paper-level extraction of sample-specific composition, processing, microstructure, and property data:

Data Extraction Prompt

```
You are an expert Materials Scientist and Data Extraction
Specialist.
Read the following excerpts from a materials science paper
regarding
Silicon Nitride (Si3N4) ceramics.  The excerpts include main
text,
tables, and figure captions.

[PAPER CONTENT]:
--- Main Text ---
{main_text}
--- Tables ---
{table_info}
--- Figure Captions ---
{figure_info}

[EXTRACTION TASK]:
Materials science papers typically compare multiple samples
(e.g., a baseline pure Si3N4 sample, and various samples with
different doping ratios or sintering parameters).  Your task
is to exhaustively extract the attributes for EVERY SINGLE
DISTINCT SAMPLE mentioned in the text and tables.
```

```
[CRITICAL RULES]:
1. Cross-Reference: You must link data from tables, text, and
captions. If Table 1 lists "Sample A" properties, and the Main
Text mentions "Sample A was sintered at 1600°C", combine these
into the SAME sample object.
2. Exhaustive: Do not miss any samples. If the paper tests 5
different compositions, output 5 distinct objects in the JSON
array.
3. Missing Data: If a specific attribute is NOT mentioned in
the provided text for a particular sample, you MUST output the
exact string "NULL". Do not leave it blank, do not guess, do
not write "N/A" or "Not provided".
4. Units: Always keep the original numerical values and units
(e.g., "1500 °C", "2 MPa", "5 wt%").
5. Output Format: Strictly output a valid JSON object
containing a single key "samples", which is an array of objects
matching the schema below.
6. Porosity vs. Relative Density: If the paper does
not explicitly provide "porosity" but provides "relative
density", you MUST extract the relative density value to
fill the porosity field. Please specify it clearly (e.g.,
"98.5% relative density") or calculate the porosity if
mathematically obvious (e.g., "1.5% (calculated from 98.5%
relative density)").
[JSON SCHEMA TEMPLATE]:
{
  "samples": [
    {
      "sample_id": "Exact name/ID of the sample (e.g., 'Sample
A',
     'Pure Si3N4', '5vol% ZrO2 doped'). Create a descriptive
one if no
     explicit ID is given.",
     "formula": {
       "silicon_nitride_powder": {
         "phase_content": {
           "alpha_si3n4": "percentage of alpha phase",
           "beta_si3n4": "percentage of beta phase"
         },
         "particle_size": "Value with unit (e.g., '0.5 μm')",
         "purity": "percentage (e.g., '99.9%')"
       },
       "sintering_additives": [
         {
           "type": "Chemical formula of additive (e.g., 'Y2O3',
           'MgO', 'Al2O3')",
           "content": "Amount with unit (e.g., '5 wt%', '2
vol%')"
         }
       ],
       "pore_forming_agent": {
         "material": "Material used for pores (e.g., 'PMMA',
         'starch')",
         "method": "Fabrication method (e.g., 'particle
stacking',
         'sacrificial template', 'gel-casting')"
       }
```

```
    },
    "process": {
      "ball_milling": {
        "speed": "rpm value",
        "time": "hours"
      },
      "forming": {
        "pressure": "MPa value (e.g., '200 MPa')"
      },
      "sintering": {
        "holding_time": "Time duration (e.g., '2 hours', '120
mins')",
        "heating_rate": "Rate (e.g., '10 °C/min')",
        "sintering_atmosphere": "Gas environment (e.g., 'N2',
        'Argon')",
        "nitrogen_pressure": "Pressure value (e.g., '0.1 MPa',
        '1 atm')"
      }
    },
    "microstructure": {
      "grains": {
        "grain_size": "Value with unit (e.g., '2.5 μm')",
        "grain_aspect_ratio": "Value (e.g., '5.2')",
        "grain_distribution": "Brief description or data"
      },
      "pores": {
        "porosity": "Percentage (e.g., '45% porosity' OR '98%
        relative density')",
        "pore_size": "Value with unit",
        "pore_shape": "Description (e.g., 'spherical',
'irregular')",
        "pore_distribution": "Brief description"
      }
    },
    "material_properties": {
      "mechanical_properties": {
        "flexural_strength": "Value with unit (e.g., '800
MPa')",
        "fracture_toughness": "Value with unit (e.g.,
        '6.5 MPa·m1/2')"
      },
      "wave_transmission_properties": {
        "frequency_bands": ["Array of strings", "e.g., 'X
band',
        'Ku band'"],
        "dielectric_constant": "Numerical value (e.g.,
'3.2')",
        "dielectric_loss": "Numerical value (e.g., '0.005')",
        "transmission_efficiency": "Percentage"
      }
    }
  }
 ]
}
```

VLM Prompt The VLM baseline was evaluated in a zero-shot setting using the following prompt, without providing ground-truth attributes or segmentation masks. The physical scale associated with each image was included when available. The complete prompt was:

VLM Prompt

```
You are an expert in materials microstructure analysis.
Carefully inspect the SEM image of a polycrystalline material
and estimate the following three quantities.
The scale of this SEM image is: 1 pixel = {um_per_px} μm.
Definitions (the estimates must follow these definitions):
   1. porosity: the percentage of the total image area
      occupied by pores, in %.
   2. grain_size: the mean major-axis length of all grains, in
      μm, converted using the scale provided above.
   3. grain_aspect_ratio: the mean ratio of the major-axis
      length to the minor-axis length over all grains;
      dimensionless.
Base your estimates solely on the image and do not fabricate
information. Strictly return a JSON object in the following
format, without any additional text:
{"porosity": <value>, "grain_size": <value>,
"grain_aspect_ratio": <value>}
```

A.12 Optimization and Data Augmentation

Optimization All SGAP variants use the same initialization and data splits. Training uses AdamW with learning rate 10^{-4} and weight decay 0.01, followed by a cosine-annealing schedule. The batch size is 2 with four gradient-accumulation steps, gradient norm clipping at 1.0, and mixed precision when CUDA is available. Training runs for at most 80 epochs with early stopping patience 15 based on validation performance. Stage I updates the structural and prediction heads together with the final two ViT blocks and pixel decoder; Stage II updates only LoRA and the prediction heads.

Data Augmentation Training-time augmentation independently applies horizontal and vertical flips with probability 0.5 and uniformly samples rotations from 0, 90, 180, and 270 degrees. Brightness and contrast jitter are each applied with probability 0.5, using a factor sampled uniformly from $[0.85, 1.15]$. The same geometric transform is applied to the image, grain-instance mask, and optional second-phase mask; nearest-neighbor interpolation is used for masks. No augmentation is used for validation. The structural and attribute losses are evaluated with the validity masks described above, so records with incomplete annotations remain usable without treating missing values as zeros.

A.13 Ablation Results Data Table

Table 4 summarizes the configurations in ablation study and additionally reports SGAP without the optional Dual-Expert branch.

Table 4: **Module-ablation results on the collected ceramic SEM test set.** Continuous attributes report MAE / R^2; categorical attributes report balanced accuracy (%) / macro-F1. Bold indicates the best value, and underlining indicates the second-best value within each metric.

Configuration	Porosity MAE / R^2	Grain size MAE / R^2	Aspect ratio MAE / R^2	Morphology bAcc / F1	Size uniformity bAcc / F1
Ours w/o LoRA	15.676 / -0.215	1.348 / 0.038	2.504 / 0.106	51.64 / 0.517	55.04 / 0.519
Ours w/o Foreground-Guided	<u>6.800</u> / <u>0.704</u>	1.099 / 0.150	1.871 / **0.581**	68.52 / 0.688	61.09 / 0.602
Ours w/o Separate Structural Path	6.802 / 0.698	**1.067** / <u>0.216</u>	2.237 / 0.414	68.84 / 0.704	65.86 / 0.647
SGAP (Ours)	7.809 / 0.635	1.205 / 0.063	**1.820** / <u>0.561</u>	<u>69.73</u> / **0.735**	<u>69.97</u> / <u>0.701</u>
SGAP with Dual-Expert (Ours)	**6.666 / 0.712**	<u>1.078</u> / **0.272**	1.934 / 0.531	**70.68** / <u>0.722</u>	**71.34 / 0.705**

The table 5 presents the data from Fig. 5 clearly.

Table 5: **Effect of different attribute-supervision data-use strategies on attribute prediction.** We report MAE, RMSE, R^2, MAPE, and sMAPE; MAPE and sMAPE are percentages. Bold indicates the best value, and underlining indicates the second-best value within each attribute and metric.

Attribute	Method	MAE↓	RMSE↓	R^2 ↑	MAPE(%) ↓	sMAPE(%) ↓
Porosity	Masked-only	15.926	26.644	-0.331	278.4	117.5
	+ Unmasked Data	15.631	25.649	-0.233	350.0	115.5
	+ ViT FT	6.998	12.599	0.702	181.9	69.5
	Ours	**6.666**	**12.396**	**0.712**	**132.4**	**65.6**
Grain size	Masked-only	1.566	2.684	-0.132	103.2	77.4
	+ Unmasked Data	1.373	2.493	0.023	95.5	67.9
	+ ViT FT	1.142	2.439	0.065	**67.7**	54.6
	Ours	**1.078**	**2.153**	**0.272**	76.0	**51.3**
Aspect ratio	Masked-only	3.233	5.233	-0.533	**46.2**	65.1
	+ Unmasked Data	2.517	4.006	0.101	60.7	46.5
	+ ViT FT	2.127	2.991	0.499	58.8	40.4
	Ours	**1.934**	**2.895**	**0.531**	50.2	**39.1**

A.14 Evaluation Metrics

MAE, RMSE, MedAE, and R^2 use their standard definitions, with MedAE denoting the median absolute error. MAPE is evaluated only for nonzero targets, whereas sMAPE omits pairs for which both prediction and target are zero:

$$\mathrm{MAPE} = \frac{100}{|\mathcal{J}_{\neq 0}|} \sum_{j \in \mathcal{J}_{\neq 0}} \frac{|\widehat{y}_j - y_j|}{|y_j|}, \qquad \mathrm{sMAPE} = \frac{100}{|\mathcal{J}_{+}|} \sum_{j \in \mathcal{J}_{+}} \frac{2|\widehat{y}_j - y_j|}{|\widehat{y}_j| + |y_j|}, \tag{28}$$

where $\mathcal{J}_{\neq 0} = \{j : |y_j| > \epsilon\}$ and $\mathcal{J}_{+} = \{j : |\widehat{y}_j| + |y_j| > \epsilon\}$. Porosity and SPAF are reported in percentage points, grain size in μm, and aspect ratio as a dimensionless quantity. For categorical attributes, balanced accuracy is the mean per-class recall and macro-F1 is the unweighted mean of the classwise F1 scores. All metrics use only samples with the corresponding valid label.