SEAR: SEEING WHAT MATTERS VIA STRUCTURE-GUIDED LANGUAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained and specialized image classification often relies on subtle differences in visual structure and appearance, including shape, color, texture, and localized defect patterns. Image-level labels provide limited guidance on which cues matter and where they occur. Pretrained vision–language models can make visual attributes explicit in text, but generic captions may not highlight the structural properties that distinguish classes. We introduce **SEAR** (Semantic Enhancement via Alignment and Reconstruction Network) to make discriminative visual structure explicit in language and fuse the resulting descriptions with image features, helping identify subtle but decisive cues. A frozen VLM generates these descriptions under explicit instructions to focus on parts, shapes, and local patterns. Alignment encourages semantic consistency between the two views, while reconstruction helps preserve complementary visual evidence that the descriptions may omit. A finite-sample analysis characterizes a text-quality threshold: generated text helps when the recoded structural signal outweighs generator noise and the additional estimation cost. Across eight benchmarks, SEAR improves mean F1-macro over the ResNet-101 image-only baseline by 0.32–7.68 percentage points, including 2.43 points on Steel Defect, where text alone performs poorly. On a 20-class CUB subset, SEAR with task-prompted descriptions also outperforms its counterpart using human-authored captions, suggesting that faithfully describing image content may not emphasize the attributes most useful for classification.

1 INTRODUCTION

Fine-grained and specialized image classification often depends on subtle structural cues, such as a narrow scratch on a steel surface, the distribution of spots on a leaf, or the ear shape of a pet breed. Modern visual backbones (He et al., 2016; Dosovitskiy et al., 2021; Liu et al., 2022; Zhu et al., 2024) encode these cues alongside variations in lighting, background, and overall appearance, and image-level labels indicate neither which structural properties distinguish classes nor where they occur.

Language can make such cues explicit. Prior work uses text to enrich category representations (Zhou et al., 2022b;a; Menon & Vondrick, 2022; Pratt et al., 2023; Roth et al., 2023; Khattak et al., 2024), to guide visual representation learning during training (Lewis et al., 2023; Nauen et al., 2026), or to fuse image features with descriptions from a captioner trained on human-authored text (Nishida et al., 2022). Two questions remain: how and when re-expressing visual evidence in language exposes class-relevant cues that an image-only classifier underuses, and how such potentially incomplete or inaccurate descriptions should be fused with image features.

We introduce **SEAR** (Semantic Enhancement via Alignment and Reconstruction Network), which makes class-relevant visual structure explicit in language while preserving complementary visual evidence. A structural prompt encodes task understanding and guides a frozen VLM to describe parts, shapes, and local patterns without specifying class decision rules. The resulting descriptions recode visual evidence into explicit linguistic attributes and relations that the classifier can associate with class labels (Figure 1). Because they may contain errors or omit details, SEAR uses alignment to emphasize attributes consistently expressed in image and text, and self-reconstruction of encoder features to retain information that alignment may suppress.

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.

An accurate caption need not emphasize the structures that distinguish classes: in leaf-disease classification, 'a green leaf on a branch" conveys image content, whereas 'small brown spots clustered along the leaf margin" makes spot appearance and distribution explicit. On a 20-class CUB subset with five descriptions per image, SEAR achieves higher F1-macro with structure-guided descriptions than with human-authored ones (Figure 2).

Our finite-sample analysis decomposes excess risk into representation bias and estimation error. Structure-guided descriptions reduce bias when they cover target-relevant structure that the image representation does not fully capture, and semantic recoding can offset the estimation cost of the added coordinates when the fused regression is sufficiently smooth. Our experiments support two implications. First, the descriptions need not classify well on their own: on Steel Defect, generated text alone reaches only 42.63 F1-macro, yet SEAR improves on the image-only baseline by 2.43 points. Second, the gain depends on description quality: with a 3B instead of a 7B generator, SEAR falls back to the image-only range.

Across eight benchmarks spanning industrial inspection, remote sensing, agriculture, and medical and natural images, SEAR improves mean F1-macro over the image-only baseline on all eight, by 0.32–7.68 points, whereas alignment-only and reconstruction-only variants each fail to exceed it on three. Our contributions are:

- **SEAR**, which fuses structure-guided descriptions from a frozen VLM with image features through alignment and self-reconstruction (Section 3).
- A finite-sample analysis of when such descriptions reduce prediction risk (Section 4).
- Experiments on eight benchmarks with diagnostics on fusion objectives, generator quality, and description source (Section 5).

2 Related Work

Language-assisted image classification. Fine-grained recognition depends on subtle attributes and local cues (Wei et al., 2022), which natural-language descriptions can convey (Reed et al., 2016; He & Peng, 2017). At the category level, prompt-learning methods adapt textual category representations (Zhou et al., 2022b;a; Khattak et al., 2024), while descriptor-based methods enrich them with generated descriptions or prompt ensembles (Menon & Vondrick, 2022; Pratt et al., 2023; Roth et al., 2023; Saha et al., 2024). At the image level, LIDE trains a captioner on human-authored descriptions and fuses image and text features during both training and inference (Nishida et al., 2022). These studies establish the value of language for image classification. The key question is whether generated descriptions make underused class-relevant visual structure accessible to the classifier, rather than add redundancy or noise.

Generated data and training-time semantic guidance. Generative augmentation expands training sets with synthesized images (Trabucco et al., 2023; Lingenberg et al., 2024). Language can also guide visual representation learning while keeping inference image-only: GIST matches generated descriptions to training images within each class (Lewis et al., 2023), and TextTeacher uses caption embeddings as auxiliary supervision (Nauen et al., 2026). Related methods use CLIP- or VLM-based distillation and pseudo-labeling to transfer semantic guidance into image-only classifiers (Li et al., 2023b; Abdelfattah et al., 2023; Kim & Shim, 2025). Missing-modality methods, designed for auxiliary modalities that may be absent at test time, likewise deploy predictors that act on image features alone (Ma et al., 2021; Zuo et al., 2022). In the reverse direction, LUGPI uses generated images to support text classification by distilling a multimodal teacher into a text-only student (Menadil et al., 2023). SEAR instead retains the generated textual view during both training and inference, so each prediction uses image-specific semantic cues alongside visual features.

Vision–language models as semantic generators. Vision–language pretraining supports transferable image–text representations (Radford et al., 2021; Jia et al., 2021) and image-conditioned generation (Li et al., 2022; 2023a), while instruction prompts (Liu et al., 2023; Bai et al., 2025) and multimodal in-context examples (Alayrac et al., 2022) steer what is described. Yet accurate descriptions need not make class-relevant structure explicit, and generated text may contain visually unsupported details (Rohrbach et al., 2018; Li et al., 2023c).

Alignment and reconstruction for multi-view learning. Multi-view learning seeks agreement across views while retaining complementary information: Deep CCA learns correlated nonlinear representations (Andrew et al., 2013), DCCAE adds reconstruction (Wang et al., 2015), MISA combines shared and modality-specific representations with alignment and reconstruction (Hazarika et al., 2020), and DMVAE models shared and private latent variables (Lee & Pavlovic, 2021). When the textual view is generated from the image rather than supplied by the dataset, the challenge is to maintain semantic consistency while preserving visual evidence that descriptions may omit or misrepresent.

SEAR addresses these issues with structural prompts that guide a frozen pretrained VLM to describe parts, shapes, and local patterns, without training a task-specific captioner. It combines alignment with self-reconstruction of encoder features to reinforce semantic consistency and preserve complementary visual evidence for classification. Its finite-sample analysis characterizes when such descriptions reduce representation bias and when semantic recoding offsets the added estimation cost.

3 The SEAR Framework

SEAR generates a textual view of each image with a frozen vision–language model (VLM) and learns to classify from the resulting image–text pair. Section 3.1 defines the two branches and their representations; Section 3.2 describes how the textual view is generated; and Section 3.3 presents the joint training objective and inference procedure.

3.1 Inputs and Representations

We consider supervised image classification with a dataset $\mathcal{D} = \{(x_i^t, y_i)\}_{i=1}^n$, where x_i^t is an image and y_i its label. A frozen pretrained VLM $\mathcal{G}$, with weights ω and a structural prompt ρ, generates the corresponding textual view $x_{i,\mathcal{G}}^s := \mathcal{G}(x_i^t; \omega, \rho)$. The superscripts t and s denote the image and textual views, respectively; the subscript $\mathcal{G}$ distinguishes generated text from an independently observed view. With ω and ρ fixed, the textual view is generated from each query image and used during both training and inference.

A frozen image encoder E_t and a frozen generated-text encoder E_s produce encoder embeddings $\hat{e}$. Trainable projection heads p_t and p_s map these embeddings into a common d-dimensional space to obtain structured representations $\hat{z}$: $\hat{e}_i^t = E_t(x_i^t)$, $\hat{e}_{i,\mathcal{G}}^s = E_s(x_{i,\mathcal{G}}^s)$, $\hat{z}_i^t = p_t(\hat{e}_i^t) \in \mathbb{R}^d$, and $\hat{z}_{i,\mathcal{G}}^s = p_s(\hat{e}_{i,\mathcal{G}}^s) \in \mathbb{R}^d$. The projection heads use separate weights because the two views express visual structure differently. The generator names and organizes discriminative attributes in text, whereas the image embeddings retain these cues alongside pose, lighting, and background variations. Separate heads allow each branch to extract the relevant structure from its own encoder embeddings. We write $\phi_t = p_t \circ E_t$ and $\tilde{\phi}_s = p_s \circ E_s$, so that $\hat{z}_i^t = \phi_t(x_i^t)$ and $\hat{z}_{i,\mathcal{G}}^s = \tilde{\phi}_s(x_{i,\mathcal{G}}^s)$. The fusion predictor uses the pair $(\hat{z}_i^t, \hat{z}_{i,\mathcal{G}}^s)$ for classification (Section 3.3).

Text generation is performed offline. The generator and both encoders remain frozen during downstream training; Section 3.3 specifies the trainable components. Encoder and projection dimensions are given in Appendix D.1.

3.2 Structure-Guided Text Generation

Knowledge injection. Raw pixel values do not explicitly express abstract structure. Human cognition draws on conceptual knowledge and an understanding of the problem to extract key structural information from an image and filter out irrelevant content (Castelhano et al., 2009). SEAR combines human task understanding, expressed through the structural prompt ρ, with the visual–language knowledge encoded in the frozen VLM's pretrained weights ω. This knowledge enables the generator to recognize and describe visual structure; the prompt guides it to retain class-relevant parts, shapes, and local patterns while filtering out irrelevant content.

The prompt $\rho = (\rho_{\text{instr}}, \mathcal{S})$ combines a structural instruction with optional image–text examples:

1. **An explicit structural instruction.** The instruction ρ_{instr} directs $\mathcal{G}$ to describe task-relevant structure rather than provide a general scene caption. For example, the Steel Defect instruction

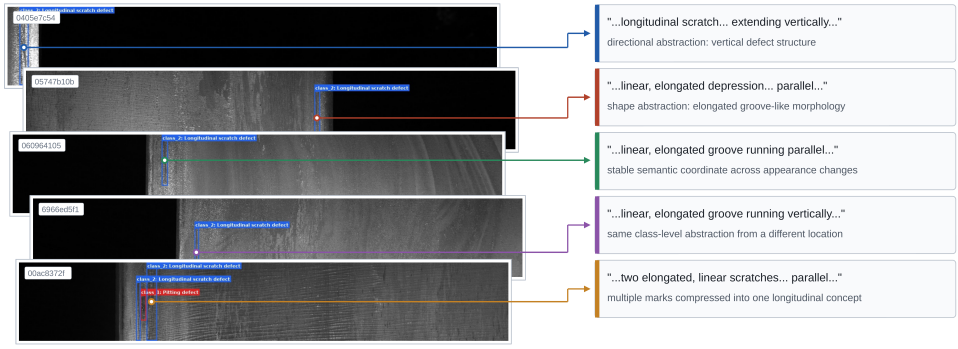


Figure 1: Longitudinal-scratch instances vary in location, texture, and local appearance, but their generated descriptions map them to stable linguistic abstractions such as longitudinal, linear, elongated, parallel, and groove.

asks the generator to describe "… surface tone, contrast, visible texture, and any visible defect characteristics, including appearance, formation cause, and general assessment of the surface condition …". The Bone Fracture instruction asks it to describe "… bone structure, any visible fractures, cracks, or abnormalities, and the overall condition of the bone …". Full prompts appear in Table 8.

2. **Few-shot image–text examples.** We include K image–text pairs $\mathcal{S} = \{(x_k^t, s_k)\}_{k=1}^K$ in the prompt context. Each description s_k names structural attributes of its paired image x_k^t. These examples demonstrate the intended descriptive focus; they neither act as a classifier nor specify class decision rules.

The examples supplement the instruction where domain-specific structure requires more concrete guidance. In our benchmarks, they are used only for Steel Defect ($K = 7$); all other datasets use the instruction alone ($\mathcal{S} = \emptyset$). Each query image requires one frozen-VLM call to produce this textual view. The query label is never provided, and $\mathcal{G}$ is not fine-tuned.

Semantic recoding expresses the selected structure in language. Figure 1 shows visually different longitudinal-scratch images described using stable terms such as longitudinal, linear, and groove. These consistent descriptions make class-relevant structure accessible to a compact downstream predictor. At matched target-approximation accuracy, the complexity comparison concerns the metric entropy of downstream prediction classes, rather than the size of the linguistic space. Section 4 analyzes how the resulting representations affect prediction bias and estimation error.

Generator residual captures image noise, omissions, hallucinated attributes, and unfocused phrasing relative to the ideal structural description. Alignment and self-reconstruction guide how the learned representations use this generated textual view, as described next in Section 3.3.

3.3 Training and Inference

SEAR jointly optimizes classification, alignment, and self-reconstruction :

$$\mathcal{L}_{\text{SEAR}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \tag{1}$$

Here $\mathcal{L}_{\text{cls}}$ is the classification loss, while λ_{align} and λ_{rec} weight alignment and self-reconstruction, respectively. The three terms are defined below.

Classification loss. The lightweight fusion predictor maps the two structured representations to logits, $\hat{\ell}_i = f_\psi(\hat{z}_i^t, \hat{z}_{i,\mathcal{G}}^s) \in \mathbb{R}^C$. Here C is the number of classes and $\hat{\ell}_{i,c}$ is the logit for class c. For a

batch of B multi-label examples, the classification loss is binary cross-entropy with logits:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{BC}\sum_{i=1}^{B}\sum_{c=1}^{C}\big[y_{i,c}\log\sigma(\hat{\ell}_{i,c}) + (1-y_{i,c})\log(1-\sigma(\hat{\ell}_{i,c}))\big], \tag{2}$$

Single-label tasks use standard cross-entropy. In both cases, classification supervises the fused logits and propagates label information into both branches.

Alignment loss. Alignment uses cross-view agreement to emphasize image-supported descriptive content and reduce the influence of generator errors. The generated view can omit visual evidence, introduce hallucinated attributes, or use generic phrasing, whereas attributes grounded in the paired image provide a basis for semantic consistency. We implement alignment through a contrastive loss between $\hat{z}^t$ and $\hat{z}^s_{\mathcal{G}}$. Let $\bar{z}^t_i = \hat{z}^t_i/\|\hat{z}^t_i\|_2$ and $\bar{z}^s_{i,\mathcal{G}} = \hat{z}^s_{i,\mathcal{G}}/\|\hat{z}^s_{i,\mathcal{G}}\|_2$, and define the cross-view similarity $s_{i,j} = (\bar{z}^t_i)^\top \bar{z}^s_{j,\mathcal{G}}/\tau$ with temperature $\tau > 0$. The symmetric InfoNCE loss is

$$\mathcal{L}_{\text{align}} = \tfrac{1}{2}\big(\mathcal{L}_{t\to s} + \mathcal{L}_{s\to t}\big), \quad \mathcal{L}_{t\to s} = -\frac{1}{B}\sum_i \log\frac{\exp(s_{i,i})}{\sum_j \exp(s_{i,j})}, \quad \mathcal{L}_{s\to t} = -\frac{1}{B}\sum_i \log\frac{\exp(s_{i,i})}{\sum_j \exp(s_{j,i})}. \tag{3}$$

The positive pair $(\hat{z}^t_i, \hat{z}^s_{i,\mathcal{G}})$ comes from the same instance; negatives are mismatched pairs within the batch. The objective encourages agreement on attributes, local patterns, and contextual relations visible in the image and verbalized in the description, favoring image-supported content over unsupported descriptions. Generic phrasing shared across images contributes little to distinguishing matched from mismatched pairs. With a weak generator, however, alignment can pull the image branch toward noisy or partly hallucinated text representations at the expense of reliable visual evidence.

Self-reconstruction loss. Self-reconstruction counters the loss of evidence that alignment may suppress, including localized visual cues absent from a sparse or inaccurate description. Each branch uses its own reconstruction head, g^{rec}_t or g^{rec}_s, to recover its encoder embedding $\hat{e}$ from its structured representation $\hat{z}$:

$$\tilde{e}^t_i = g^{\text{rec}}_t(\hat{z}^t_i), \qquad \tilde{e}^s_{i,\mathcal{G}} = g^{\text{rec}}_s(\hat{z}^s_{i,\mathcal{G}}), \qquad \mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_i \Big(\|\hat{e}^t_i - \tilde{e}^t_i\|_2^2 + \|\hat{e}^s_{i,\mathcal{G}} - \tilde{e}^s_{i,\mathcal{G}}\|_2^2\Big). \tag{4}$$

Reconstruction operates within each branch: each structured representation is required to retain the information needed to recover its own encoder embedding, whether or not the other view expresses that content. In the image branch, this directly penalizes the loss of visual evidence when the generated description is sparse or generic. In the text branch, it encourages retention of description-specific structure beyond the content shared by both views.

The reconstruction target is the fixed encoder embedding $\hat{e}$. Classification and alignment shape the projected representation $\hat{z}$ for prediction and cross-view agreement. Reconstructing $\hat{z}$ would therefore tie the target to the same pressures that can suppress complementary evidence. Reconstructing $\hat{e}$ instead anchors each branch to its encoder-level evidence before projection. Table 2(b) compares the two targets empirically.

Joint optimization. The three losses have complementary and interdependent roles: classification links both branches to the labels, alignment promotes semantic consistency, and self-reconstruction anchors each branch to its encoder-level evidence. Alignment rewards cross-view correspondence and self-reconstruction rewards within-view fidelity; neither alone requires class-discriminative representations. Classification makes both serve the prediction task.

We jointly train the projection heads, fusion predictor, and reconstruction heads, with $\theta = \{p_t, p_s, f_\psi, g^{\text{rec}}_t, g^{\text{rec}}_s\}$. The generator $\mathcal{G}$ and encoders E_t and E_s remain frozen and are excluded from θ.

Inference. At inference, $\mathcal{G}$ generates the textual view of the query image. The two branches produce the structured representations defined in Section 3.1, and the fusion predictor computes logits from their pair. The reconstruction heads are used only during training. The final prediction is

$$\hat{y}_i = \text{Dec}(\hat{\ell}_i), \qquad \text{Dec}(\hat{\ell}_i) = \begin{cases} \arg\max_c \hat{\ell}_{i,c} & \text{single-label,} \\ \big(\mathbf{1}\{\sigma(\hat{\ell}_{i,c}) > \eta\}\big)_{c=1}^{C} & \text{multi-label,} \end{cases} \tag{5}$$

with the multi-label threshold η tuned on the validation set.

4 When and Why the Generated View Helps

This section explains how human-guided structural descriptions improve prediction from frozen representations. Section 4.1 formalizes the structural mechanism in Section 3.2 and defines the regression setting. Section 4.2 shows how target-relevant structural coverage reduces representation bias, and Section 4.3 examines how regression smoothness offsets the estimation cost of added coordinates. Together, these results identify conditions for a reduction in prediction risk.

4.1 Structural Representation and Regression Setup

Let u and v^t denote shared semantic factors and image-specific factors, respectively, both associated with the target y. The image is modeled as $x^t = g_t(u, v^t) + \epsilon_t$, where ϵ_t denotes observation noise. The textual view is modeled as $x^s = g_s(u, v^s) + \epsilon_s$, where v^s denotes private structural information expressed in text, and ϵ_s captures background details, stylistic variation, and other content unrelated to classification.

For the subsequent squared-loss analysis, we model the target as $y = g(u, v^t, v^s) + \varepsilon$, where g describes the target's dependence on the visual factors (u, v^t) and the structural information v^s, and ε denotes label noise.

Since v^s is independent of (u, v^t, ϵ_t), it could not be recovered from the observed image alone. For generated views, v^s denotes task-relevant structural attributes that are present in the image but not fully captured by the frozen image representation. To construct a textual view from the available visual evidence, we introduce the mapping $q_{\omega,\rho}(u, v^t)$ following Section 3.2. This mapping combines pretrained visual–language knowledge encoded in ω with human conceptual knowledge and task understanding conveyed through ρ to identify task-relevant structural attributes and relations while filtering out irrelevant content. The renderer $\tilde{g}$ then expresses the selected structure in language. Accordingly, the generated representation can be written as $\hat{e}^s_{\mathcal{G}} = E_s(\tilde{g}(q_{\omega,\rho}(u, v^t))) + \epsilon_{\mathcal{G}}$. The image supplies instance-specific content, while (ω, ρ) guide the selection, organization, and linguistic expression of task-relevant structural features. These features are precisely what v^s represents.

4.2 Excess Risk and Structural Coverage

We compare prediction from the frozen image representation $\hat{e}^t$ with prediction from the fused representation $(\hat{e}^t, \hat{e}^s_{\mathcal{G}})$. Let $\mathcal{F}_{\text{img}}$ denote the class of predictors with finite second moment that use only $\hat{e}^t$, and let $\mathcal{F}_{\text{fused}}$ denote the corresponding class using $(\hat{e}^t, \hat{e}^s_{\mathcal{G}})$. Their representation-level conditional predictors are $\tilde{f}_{\text{img}}(\hat{e}^t) = \mathbb{E}\left[g(u, v^t, v^s) \mid \hat{e}^t\right]$, and $\tilde{f}_{\text{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}}) = \mathbb{E}\left[g(u, v^t, v^s) \mid \hat{e}^t, \hat{e}^s_{\mathcal{G}}\right]$.

The corresponding representation biases are

$$\text{Bias}(\mathcal{F}_{\text{img}})^2 = \mathbb{E}\left[\left(g(u, v^t, v^s) - \tilde{f}_{\text{img}}(\hat{e}^t)\right)^2\right], \tag{6}$$

$$\text{Bias}(\mathcal{F}_{\text{fused}})^2 = \mathbb{E}\left[\left(g(u, v^t, v^s) - \tilde{f}_{\text{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}})\right)^2\right]. \tag{7}$$

For a predictor $\hat{f}_n \in \mathcal{F}$ learned from n training samples, define the excess risk above the label-noise variance as $B_n(\mathcal{F}) := \mathbb{E}[(y - \hat{f}_n)^2] - \mathbb{E}[\varepsilon^2]$, and the estimation error as $\mathcal{E}_n(\mathcal{F}) = \mathbb{E}[(y - \hat{f}_n)^2] - \inf_{f \in \mathcal{F}} \mathbb{E}[(y - f)^2]$. Expectations average over the independent training samples and an independent test pair.

Assumption 1 (Regression regularity)**.** *The regression function and label noise have finite second moments,* $\mathbb{E}\left[g(u, v^t, v^s)^2\right] < \infty, \qquad \mathbb{E}[\varepsilon^2] < \infty$*, and* $\mathbb{E}[\varepsilon \mid u, v^t, v^s, x^t, x^s_{\mathcal{G}}] = 0$.

Here, $x^s_{\mathcal{G}} = \mathcal{G}(x^t; \omega, \rho)$ is the generated textual view defined in Section 3.1.

Assumption 2 (Structural coverage and target relevance)**.** *The generated representation retains structural information that is not fully captured by the image representation:* $\hat{e}^s_{\mathcal{G}} \not\perp v^s \mid \hat{e}^t$.

The retained structure is target-relevant: the generated representation explains part of the image-only regression residual,

$$\mathbb{E}\Big[g(u, v^t, v^s) - \tilde{f}_{\text{img}}(\hat{e}^t) \mid \hat{e}^t, \hat{e}^s_{\mathcal{G}}\Big] \neq 0 \tag{8}$$

on a set of positive probability.

Under Assumption 1, the excess risk decomposes as

$$B_n(\mathcal{F}) = \mathrm{Bias}(\mathcal{F})^2 + \mathcal{E}_n(\mathcal{F}), \tag{9}$$

where

$$\mathrm{Bias}(\mathcal{F})^2 = \inf_{f \in \mathcal{F}} \mathbb{E}[(y - f)^2] - \mathbb{E}[\varepsilon^2]. \tag{10}$$

Theorem 1 (Bias reduction through human-guided structure)**.** *Under Assumption 1,*

$$\begin{aligned} \Delta &:= \mathrm{Bias}(\mathcal{F}_{\mathrm{img}})^2 - \mathrm{Bias}(\mathcal{F}_{\mathrm{fused}})^2 \\ &= \mathbb{E}\left[\left(\tilde{f}_{\mathrm{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}}) - \tilde{f}_{\mathrm{img}}(\hat{e}^t)\right)^2\right] \geq 0. \end{aligned} \tag{11}$$

Under Assumptions 1 and 2, the generated structural information yields a strict bias reduction: $\Delta > 0$.

Theorem 1 shows that generated structure reduces representation bias when it explains target variation left unresolved by the frozen image representation. The gain Δ measures this contribution; its effect on finite-sample prediction also depends on estimation error.

4.3 Estimation Rates and Risk Improvement

Structural coverage determines the reduction in representation bias. We next examine how semantic recoding affects estimation through the dimension and smoothness of the resulting regression.

Assumption 3 (Dimension and smoothness of the two regressions)**.** *Using fixed regression coordinates obtained from the respective representations, the image-only regression family ranges over a Hölder ball of order $\beta > 0$ in dimension $d_0 \geq 1$. The fused family has dimension $d_0 + k$, with $k \geq 1$, and ranges over either a linear regression family or a Hölder ball of order $\beta' > \beta$. The families satisfy the standard design and noise regularity conditions for the minimax rates in Stone (1982); Mourtada (2022), with fixed dimensions and regularity constants.*

Here, d_0 and $d_0 + k$ are regression dimensions, distinct from the embedding dimensions and the projection dimension d in Section 3.1. The additional dimension k and the image-only dimension d_0 remain fixed as n grows.

Theorem 2 (Estimation gains from smoother structural representations)**.** *Under Assumption 3, denote the minimax estimation risks of the corresponding regression families by $r^\star_{\mathrm{img}}(n)$ and $r^\star_{\mathrm{fused}}(n)$. The image-only family satisfies $r^\star_{\mathrm{img}}(n) \asymp n^{-2\beta/(2\beta+d_0)}$.*

For the fused Hölder family, $r^\star_{\mathrm{fused}}(n) \asymp n^{-2\beta'/(2\beta'+d_0+k)}$.

For the fused linear family, $r^\star_{\mathrm{fused}}(n) \asymp \frac{d_0+k}{n}$.

The fused Hölder family has a faster rate when

$$\frac{\beta'}{2\beta' + d_0 + k} > \frac{\beta}{2\beta + d_0}, \qquad \textit{equivalently} \qquad \beta' d_0 > \beta(d_0 + k).$$

The two results explain complementary benefits of human-guided structure. The image provides sample-specific content, while the knowledge and task understanding expressed through ρ organize it into structural attributes v^s. Under target-relevant coverage, these attributes reduce representation bias (Theorem 1). When semantic recoding also yields a smoother or linear regression, it can improve estimation rates (Theorem 2).

For the fitted predictors, let $r_f(n) = \mathcal{E}_n(\mathcal{F}_{\mathrm{fused}}) - \mathcal{E}_n(\mathcal{F}_{\mathrm{img}})$ denote the estimation-error difference. The risk decomposition gives

$$B_n(\mathcal{F}_{\mathrm{img}}) - B_n(\mathcal{F}_{\mathrm{fused}}) = \Delta - r_f(n). \tag{12}$$

Thus the generated view improves squared prediction risk whenever the structural bias reduction outweighs the additional estimation error, $\Delta > r_f(n)$.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on eight image classification benchmarks spanning industrial inspection, remote sensing, agriculture, natural images, fine-grained pet recognition, and medical imaging; dataset sizes and splits are summarized in Table 4 in Appendix C. Unless otherwise noted, all experiments use the same default configuration: a frozen ResNet-101 image encoder E_t, a frozen Qwen2.5-VL-7B-Instruct model for offline text generation and text encoding, and F1-macro as the primary metric. Hyperparameters are tuned per dataset with Optuna; implementation details, prompts, and selected hyperparameters are provided in Appendices D.1, E, and G. Unless otherwise stated, entries reported with $\pm$ are means and standard deviations over three random seeds.

The main table compares IMAGE-ONLY and TEXT-ONLY baselines with four generated-view fusion variants. The two baselines use image features and generated-text embeddings, respectively. SEAR-BASE uses the classification loss alone; SEAR-A and SEAR-R add alignment and self-reconstruction, respectively; and full SEAR combines both auxiliary terms in Eq. (1). Additional image-backbone and rule-based text controls are reported in Appendix D.7.

5.2 MAIN RESULTS

Table 1: F1-macro (%) across eight datasets, ordered by standalone generated-text performance. SEAR-A and SEAR-R are objective diagnostics.

Dataset	IMAGE-ONLY	TEXT-ONLY	SEAR-BASE	SEAR-A	SEAR-R	SEAR
Steel Defect	85.72±0.60	42.63±0.49	86.31±0.62	87.56±0.84	87.61±0.84	**88.15±0.57**
EuroSAT	96.67±0.38	48.06±1.23	96.18±0.32	96.48±0.16	96.39±0.19	**96.99±0.34**
Blood Cell	98.73±0.19	64.11±0.71	97.63±1.82	**99.53±0.19**	99.13±0.38	99.07±0.34
Bone Fracture	91.61±0.16	68.58±1.11	91.26±1.17	90.58±0.93	91.16±0.57	**92.08±0.76**
Pneumonia	94.25±0.70	70.17±1.14	95.98±0.62	**96.24±0.36**	95.73±0.60	95.66±0.23
Beans	94.84±1.85	80.55±1.47	96.00±1.45	96.25±0.80	94.83±0.69	**96.29±0.83**
Oxford Pets	94.43±0.77	89.03±0.74	94.52±1.47	94.37±1.06	94.73±0.62	**95.72±0.62**
Cats vs Dogs	91.65±1.72	97.66±1.89	98.67±1.25	**99.67±0.47**	98.33±0.94	99.33±0.47
Exceeds IMAGE-ONLY	–	1/8	5/8	5/8	5/8	8/8

Table 1 orders the datasets by the F1-macro of TEXT-ONLY, an empirical indicator of generated-text quality for classification. SEAR exceeds IMAGE-ONLY on all eight datasets. The benefit is particularly informative when generated text is a weak standalone predictor. On Steel Defect, the prompt also asks the generator to name the defect type, yet TEXT-ONLY reaches only 42.63 and the text branch within SEAR reaches 33.25 (Table 2(a)). The 2.43-point fusion gain supports the value of structural descriptions beyond the generator's categorical judgment. A weak standalone predictor can still provide a useful view: under Section 4.2, generated attributes v^s contribute when they help explain target variation left unresolved by the frozen image representation. Appendix D.2 discusses the strong-text end of the ordering and the limits of TEXT-ONLY as an indicator of this contribution.

Unregularized fusion (SEAR-BASE) falls below IMAGE-ONLY only on EuroSAT, Blood Cell, and Bone Fracture, all among the five datasets with TEXT-ONLY F1-macro below 71. SEAR improves on SEAR-BASE on seven of eight datasets; on Pneumonia, the two are within one standard deviation. The auxiliary objectives provide complementary benefits: SEAR-A and SEAR-R each fail to exceed IMAGE-ONLY on three datasets, including EuroSAT and Bone Fracture, while full SEAR is the only variant above it on all eight. On Bone Fracture, alignment alone reaches 90.58, consistent with the mechanism in Section 3.3: encouraging agreement with an unreliable description can suppress useful visual evidence. Adding self-reconstruction restores the gain (92.08). Alignment alone is slightly stronger on Pneumonia, Cats vs Dogs, and Blood Cell, while the full objective provides the most consistent improvement over the image-only reference across the eight datasets.

5.3 ABLATION AND DIAGNOSTIC STUDIES

Table 2 examines branch contributions, reconstruction targets, and generator capacity on Steel Defect. Panel (a) evaluates single-branch heads on the projected SEAR representations; these probes are distinct from the separately tuned IMAGE-ONLY system. The fused representation outperforms either branch. This supports semantic recoding through complementary information in the combined representation.

Table 2: Diagnostic experiments on Steel Defect. All results are F1-macro (%) unless otherwise noted.

Setting	F1-macro
(a) Branch contribution	
$\hat{z}^t$ only (image)	$85.03_{\pm 1.19}$
$\hat{z}^s_G$ only (generated text)	$33.25_{\pm 1.44}$
$[\hat{z}^t;\ \hat{z}^s_G]$ (full)	$\mathbf{88.15_{\pm 0.57}}$
(b) Reconstruction target	
Encoder embeddings (proposed)	$\mathbf{88.15_{\pm 0.57}}$
Learned representations ($\hat{z}$)	$86.26_{\pm 1.26}$
(c) Generator capacity	
IMAGE-ONLY reference	$85.72_{\pm 0.60}$
Qwen2.5-VL-7B / default	$\mathbf{88.15_{\pm 0.57}}$
Qwen2.5-VL-3B	$85.45_{\pm 1.03}$

Panel (b) shows that frozen encoder embeddings provide a more effective reconstruction target than the post-projection structured representations $\hat{z}$, labeled "learned representations" in the table. Reconstructing the encoder embeddings therefore helps preserve information that the projected representations may omit (Section 3.3).

Panel (c) replaces the 7B generator with its 3B counterpart while keeping the prompt, architecture, and training protocol fixed. With 3B text, performance returns to the image-only range, whereas 7B text preserves the gain. The regularized objective does not substitute for informative generated text (Section 4.2).

5.4 ALTERNATIVE BASELINES ON STEEL DEFECT

Table 3 compares language-use and localization baselines on Steel Defect under their respective input settings. The detailed version with backbone and inference input is given in Table 5 in Appendix D.6. The low CLIP label-match score indicates that class names alone provide limited discrimination in this industrial setting. LIDE-style captioner training (Nishida et al., 2022) requires human-written descriptions, which none of the eight benchmarks provides; using a frozen generator removes this requirement.

Table 3: Steel Defect comparison with training-time language-use methods, class-name prompt and matching baselines, a detection-based image baseline, and SEAR. Results are F1-macro (%).

Method	F1-macro
Training-time or missing-modality methods with image-only inference	
CCD-style distillation (Kim & Shim, 2025)	$73.34_{\pm 0.95}$
Large-VLM-Distillation-OOD (Li et al., 2023b)	$73.96_{\pm 0.19}$
CDUL (Abdelfattah et al., 2023)	$72.60_{\pm 0.68}$
SMIL (Ma et al., 2021)	$73.44_{\pm 0.73}$
IF-MMIN (Zuo et al., 2022)	$72.58_{\pm 0.48}$
Class-name prompt and matching baselines	
CoOP (Zhou et al., 2022b)	$62.34_{\pm 0.64}$
CoCoOp (Zhou et al., 2022a)	$71.53_{\pm 1.01}$
CLIP label match (Radford et al., 2021)	$12.78_{\pm 0.00}$
Detection-based image baseline	
YOLO11n detection → image-level classification	$84.52_{\pm 0.71}$
Retained generated-view model	
SEAR	$\mathbf{88.15_{\pm 0.57}}$

YOLO11n uses defect localization annotations and aggregates detections into image-level predictions, yet remains below SEAR, which uses image-level labels and generated descriptions. Together with the representation probe in Table 2, this comparison supports the practical value of retaining the recomputable generated view at inference. Additional fusion methods originally developed for naturally observed auxiliary modalities are reported in Appendix D.8, adapted to the same image and generated-text features.

5.5 GENERATED VS. NATURAL TEXT COMPARATOR

We compare human-authored and generated text as auxiliary views to examine their differences; Appendix D.3 gives the full setup and results.

6 DISCUSSION

These results suggest that generated descriptions are most useful when they recode visual evidence into task-relevant structure rather than generic captions. Future use should account for description reliability, especially when generated text is weak or noisy.

AI USE STATEMENT

This work uses generative AI as part of the studied method: SEAR employs a frozen Qwen2.5-VL-Instruct model to generate structure-guided image descriptions that are then fused with image features for classification. This experimental use is specified in Sections 3.2 and 5.1, including the prompts, model usage, and evaluation protocol. Apart from this disclosed experimental component, generative AI tools were used only for limited manuscript wording assistance, such as improving readability, clarity, grammar, and phrasing of author-written text. We did not use generative AI tools to develop the method, design experiments, implement the system, analyze results, create figures, identify literature, or formulate theoretical claims. The authors reviewed all AI-assisted text and are responsible for the final manuscript.

ETHICS STATEMENT

This work studies language-assisted image classification using pretrained models and established public benchmarks spanning industrial inspection, remote sensing, agriculture, medical imaging, and natural-image recognition. We do not collect new human-subject data, release a new dataset, or use personally identifying or sensitive private information. For medical-image benchmarks, the experiments are retrospective benchmark evaluations and are not intended to provide clinical diagnosis or deployment-ready medical advice. Since generated descriptions can omit or misstate visual evidence, deployment should be preceded by task-specific validation and appropriate domain review.

REPRODUCIBILITY STATEMENT

Section 3 describes the SEAR framework, including textual-view generation, frozen image and text encoders, projection heads, alignment, reconstruction, and prediction objectives. Section 5 specifies the evaluated benchmarks, baselines, metrics, default model configuration, and evaluation protocol. The appendices provide dataset details and splits, implementation details, prompts, selected hyperparameters, additional controls, and diagnostic experiments, including generator-scale comparisons and generated-versus-human-description analyses. These details are intended to make the reported results reproducible from the stated public datasets, pretrained models, fixed prompts, and training/evaluation procedures. We will release the code after it has been organized and documented.

REFERENCES

Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. CDUL: CLIP-driven unsupervised learning for multi-label image classification. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2307.16634.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, Anchorage AK USA, July 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330701. URL https://dl.acm.org/doi/10.1145/3292500.3330701.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198, 2022. URL https://arxiv.org/abs/2204.14198.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1247–1255. PMLR, 2013.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan,

Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025. URL http://arxiv.org/abs/2502.13923.

Monica S. Castelhano, Michael L. Mack, and John M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):6.1–6.15, 2009. doi: 10.1167/9.3.6. URL https://doi.org/10.1167/9.3.6.

dmxtrx. bone-fracture-detection-dataset. Kaggle dataset, 2025. URL https://www.kaggle.com/datasets/dmtrrmnv/clean-bone-fracture. Version 5; accessed: 2026-06-22.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2021. URL http://arxiv.org/abs/2010.11929.

Jeremy Elson, John R. Douceur, Jon Howell, and Jared Saul. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*. Association for Computing Machinery, October 2007. URL https://www.microsoft.com/en-us/research/publication/asirra-a-captcha-that-exploits-interest-aligned-manual-image-categorization/.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131. Association for Computing Machinery, 2020. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5994–6002, 2017.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. arXiv preprint arXiv:1709.00029, 2017. URL https://arxiv.org/abs/1709.00029.

Prabhash Kumar Jha. Peripheral Blood Cell -2. Kaggle dataset, 2024. URL https://www.kaggle.com/datasets/prabhashkumarjha/wbc-dataset-1. Version 2; accessed: 2026-06-22.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley Data, V2, 2018. URL https://data.mendeley.com/datasets/rscbjbr9sj/2.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. arXiv preprint arXiv:2401.02418, 2024. URL https://arxiv.org/abs/2401.02418.

Dongseob Kim and Hyunjung Shim. Classifier-guided CLIP distillation for unsupervised multi-label classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.16873.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Mihee Lee and Vladimir Pavlovic. Private-Shared Disentangled Multimodal VAE for Learning of Latent Representations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1692–1700, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-899-4. doi: 10.1109/CVPRW53098.2021.00185. URL https://ieeexplore.ieee.org/document/9523016/.

Kathleen M. Lewis, Emily Mu, Adrian V. Dalca, and John Guttag. GIST: Generating image-specific text for fine-grained object classification. *arXiv preprint arXiv:2307.11315*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *IEEE/CVF International Conference on Computer Vision*, 2023b. URL https://arxiv.org/abs/2307.03135.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023c. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

Tobias Lingenberg, Markus Reuter, Gopika Sudhakaran, Dominik Gojny, Stefan Roth, and Simone Schaub-Meyer. DIAGen: Semantically Diverse Image Augmentation with Generative Models for Few-Shot Learning. arXiv preprint arXiv:2408.14584, 2024. URL https://arxiv.org/abs/2408.14584.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986. IEEE, 2022.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. arXiv preprint arXiv:2103.05677, 2021. URL https://arxiv.org/abs/2103.05677.

Makerere AI Lab. Bean Disease Dataset, January 2020. URL https://github.com/AI-Lab-Makerere/ibean/.

Rafael-Edy Menadil, Mariana-Iuliana Georgescu, and Radu Tudor Ionescu. Learning using generated privileged information by text-to-image diffusion models. arXiv preprint arXiv:2309.15238, 2023. URL https://arxiv.org/abs/2309.15238.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. arXiv preprint arXiv:2210.07183, 2022. URL http://arxiv.org/abs/2210.07183.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022. doi: 10.1214/22-AOS2181. URL https://doi.org/10.1214/22-AOS2181.

Tobias Christian Nauen, Stanislav Frolov, Brian Bernhard Moser, Federico Raue, Ahmed Anwar, and Andreas Dengel. TextTeacher: What can language teach about images? arXiv preprint arXiv:2605.22098, 2026. URL https://arxiv.org/abs/2605.22098.

Kosuke Nishida, Kyosuke Nishida, and Shuichi Nishioka. Improving few-shot image classification using machine- and user-generated natural language descriptions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1421–1430, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.106. URL https://aclanthology.org/2022.findings-naacl.106/.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and Dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15645–15655. IEEE, 2023. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01438. URL https://ieeexplore.ieee.org/document/10378461/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49–58, 2016.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437. URL http://aclweb.org/anthology/D18-1437.

Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. arXiv preprint arXiv:2306.07282, 2023. URL https://arxiv.org/abs/2306.07282.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting VLMs with text descriptions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17542–17552. IEEE, 2024. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01661. URL https://ieeexplore.ieee.org/document/10658340/.

Severstal and Kaggle. Severstal: Steel Defect Detection. Kaggle competition, 2019. URL https://www.kaggle.com/competitions/severstal-steel-defect-detection. Accessed: 2026-06-22.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective Data Augmentation With Diffusion Models. arXiv preprint arXiv:2302.07944, 2023. URL https://arxiv.org/abs/2302.07944.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1083–1092, 2015.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8927–8948, 2022. doi: 10.1109/TPAMI.2021.3126648.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 2018.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI Conference on Artificial Intelligence*, 2018.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01631. URL https://ieeexplore.ieee.org/document/9879913/.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-022-01653-1. URL http://arxiv.org/abs/2109.01134.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417, 2024. URL http://arxiv.org/abs/2401.09417.

Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. arXiv preprint arXiv:2210.15359, 2022. URL https://arxiv.org/abs/2210.15359.

A Limitations

The reported descriptions follow the prompt in Section 3.2; comparisons with generic captions and an instruction-only ablation are not included. Using a capable vision–language generator during preprocessing or at deployment increases pipeline complexity. Future work can extend generated-view fusion beyond classification, develop reliability-aware fusion that adapts to estimated text quality, and refine prompts to specify task-relevant structural attributes and relations more precisely.

B Proofs and Additional Details

B.1 Proof of Theorem 1

Proof. Write $g = g(u, v^t, v^s)$. Assumption 1 gives $\mathbb{E}[\varepsilon(g - f)] = 0$ for predictors based on either representation, so

$$\mathbb{E}[(y - f)^2] = \mathbb{E}[(g - f)^2] + \mathbb{E}[\varepsilon^2].$$

Minimizing over the corresponding predictors identifies Eqs. (6)–(7) with Eq. (10) and yields Eq. (9).

Let

$$r = g(u, v^t, v^s) - \tilde{f}_{\text{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}})$$

and

$$\delta = \tilde{f}_{\text{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}}) - \tilde{f}_{\text{img}}(\hat{e}^t).$$

Then

$$g(u, v^t, v^s) - \tilde{f}_{\text{img}}(\hat{e}^t) = r + \delta.$$

Expanding the square gives

$$\mathbb{E}\left[\left(g(u, v^t, v^s) - \tilde{f}_{\text{img}}(\hat{e}^t)\right)^2\right] = \mathbb{E}[r^2] + \mathbb{E}[\delta^2] + 2\mathbb{E}[r\delta].$$

Table 4: Summary of the eight evaluation datasets. "Multi" and "Single" denote multi-label and single-label classification, respectively.

Dataset	Domain	Classes	Task	Train	Val	Test
Steel Defect	Industrial	5	Multi	10,054	1,257	1,257
EuroSAT	Remote sensing	10	Single	21,600	2,700	2,700
Blood Cell	Medical	5	Single	4,000	500	500
Bone Fracture	Medical	2	Single	1,783	223	222
Pneumonia	Medical	2	Single	4,672	585	583
Beans	Agricultural	3	Single	934	117	116
Oxford Pets	Fine-grained	10	Single	1,584	199	198
Cats vs Dogs	Natural	2	Single	800	100	100

By the definition of the conditional predictor,

$$\mathbb{E}\left[r \mid \hat{e}^t, \hat{e}^s_{\mathcal{G}}\right] = 0,$$

and hence

$$\mathbb{E}[r\delta] = 0.$$

Therefore,

$$\Delta = \mathbb{E}[\delta^2] = \mathbb{E}\left[\left(\tilde{f}_{\text{fused}}(\hat{e}^t, \hat{e}^s_{\mathcal{G}}) - \tilde{f}_{\text{img}}(\hat{e}^t)\right)^2\right] \geq 0.$$

Under Assumption 2,

$$\delta = \mathbb{E}\Big[g(u, v^t, v^s) - \tilde{f}_{\text{img}}(\hat{e}^t) \mid \hat{e}^t, \hat{e}^s_{\mathcal{G}}\Big]$$

is nonzero on a set of positive probability. Therefore

$$\mathbb{E}[\delta^2] = \Delta > 0.$$

□

B.2 Proof of Theorem 2

Proof. The rates follow from the standard squared-error minimax results for Hölder regression and regular linear regression (Stone, 1982; Mourtada, 2022). With fixed dimension, the linear rate is faster than a nonparametric Hölder rate of finite smoothness. For the two Hölder families, comparing the exponents gives

$$\beta' d_0 > \beta(d_0 + k).$$

For the fitted predictors, subtracting the two decompositions in Eq. (9) gives Eq. (12). □

C Dataset Details

The eight datasets cover both single-label and multi-label tasks and vary in how strongly classification depends on subtle, localized, or domain-specific cues. Table 4 summarizes their domains, class counts, task types, and splits.

Steel Defect is a multi-label industrial surface inspection task derived from the Severstal Steel Defect Detection benchmark (Severstal & Kaggle, 2019). The five labels include a background/no-defect indicator and surface anomaly types such as pitting, longitudinal scratch, scratch, and roll-in defects; because the anomaly labels may co-occur, we evaluate the task as multi-label. EuroSAT is a remote-sensing land-use classification dataset (Helber et al., 2017) with ten top-down land-cover categories. Blood Cell contains five microscopy cell types (Jha, 2024). Bone Fracture and Pneumonia are medical X-ray classification datasets (dmxtrx, 2025; Kermany et al., 2018). Beans focuses on leaf disease recognition (Makerere AI Lab, 2020), Oxford Pets is a ten-class fine-grained pet breed subset (Parkhi et al., 2012), and Cats vs Dogs is a natural-image binary classification task (Elson et al., 2007).

The CUB-200-2011 experiment in Section D.3 is a diagnostic comparison rather than part of the eight-dataset main table. It uses a 20-class subset of CUB-200-2011 (Wah et al., 2011) together

with the human-authored bird descriptions from Reed et al. (2016). For each image, generated descriptions are produced independently by $\mathcal{G}$ and compared with budgets of one, three, five, or ten human descriptions; when multiple descriptions are used, they are concatenated before encoding to form the textual view.

D Additional Experimental Diagnostics

D.1 Implementation Details

All main eight-dataset experiments use text views generated offline by Qwen2.5-VL-7B-Instruct Bai et al. (2025), a frozen vision–language model serving as $\mathcal{G}$ in Section 3.1. The 3B experiments in Table 2(c) and Appendix F change only the generator scale for producing the text view; downstream training protocol, projection architecture, and evaluation are kept fixed. Following Section 3.2, each image is passed to $\mathcal{G}$ with the prompt ρ of Appendix E. $\mathcal{G}$ is not shown the query label. No task-specific fine-tuning is applied to $\mathcal{G}$. The generated text is encoded by the same frozen text-encoding pipeline to produce fixed 3,584-dimensional text embeddings; the trainable projection head p_s maps these embeddings to the structured representation $\hat{z}^s_{\mathcal{G}} \in \mathbb{R}^{256}$.

On the image side, a frozen ResNet-101 He et al. (2016) extracts 2,048-dimensional image embeddings from the raw input as E_t, and the trainable projection head p_t maps them to $\hat{z}^t \in \mathbb{R}^{256}$. The fusion predictor f_ψ is a lightweight cross-attention and MLP head over the projected image and text representations. The projection heads p_t and p_s, fusion predictor f_ψ, and reconstruction heads g^{rec}_t and g^{rec}_s are trained jointly; the ResNet-101 backbone E_t, generator $\mathcal{G}$, and generated-text encoder E_s remain frozen throughout.

All models are optimized with Adam Kingma & Ba (2015) for up to 100 epochs with batch size 32. Reported entries with $\pm$ are means and standard deviations over three random seeds. Hyperparameters, including the learning rate, dropout rates, auxiliary loss weights λ_{align} and λ_{rec}, and contrastive temperature τ, are tuned independently per dataset via Optuna Akiba et al. (2019). Optimal configurations for the full model are reported in Appendix G. Reconstruction uses MSE for all datasets. For the multi-label Steel Defect task, a single shared decision threshold η is tuned on the validation set; single-label tasks use the argmax prediction.

Model variants. SEAR-Base sets $\lambda_{\text{align}} = \lambda_{\text{rec}} = 0$ in Eq. (1). SEAR-A and SEAR-R isolate $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{rec}}$ by setting $\lambda_{\text{rec}} = 0$ and $\lambda_{\text{align}} = 0$, respectively; they are diagnostic variants of the training objective in Section 3.3. The Text-only baseline in Table 1 uses frozen generated-text embeddings before SEAR projection, whereas the generated-text probe in Table 2(a) is evaluated on the projected representation $\hat{z}^s_{\mathcal{G}}$.

D.2 Additional Analysis of Main Results

At the strong-text end of the ordering in Table 1, the VLM names the animal categories in Cats vs Dogs, where keyword parsing reaches 100.00 (Table 6). The textual view thus approaches a zero-shot classifier, and fusion benefits from this signal.

Standalone text performance reflects both complementary structure and information already available in the image. Its value for fusion therefore depends on the target-relevant structure that the generated representation contributes beyond the frozen image representation (Section 4.2); a high Text-only score need not imply a large reduction in representation bias.

D.3 Generated vs. Natural Text Comparator

This study compares human-authored descriptions (Reed et al., 2016) and generated descriptions as alternative textual views of the same image, following the representation-based analysis in Section 4. With five descriptions per image, generated text reaches 84.96, exceeding every human-authored setting tested (79.23–81.49). The advan-

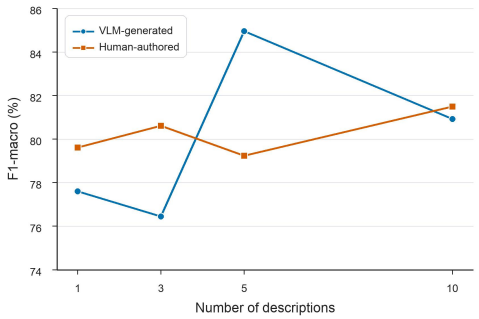


Table 5: Detailed Steel Defect comparison with training-time language-use methods, class-name prompt and matching baselines, a detection-based image baseline, and SEAR. Results are F1-macro (%).

Method	Backbone	Inference input	F1-macro
Training-time or missing-modality methods with image-only inference			
CCD-style distillation (Kim & Shim, 2025)	ResNet-101	image	$73.34_{\pm 0.95}$
Large-VLM-Distillation-OOD (Li et al., 2023b)	ResNet-101	image	$73.96_{\pm 0.19}$
CDUL (Abdelfattah et al., 2023)	ResNet-101	image	$72.60_{\pm 0.68}$
SMIL (Ma et al., 2021)	ResNet-101	image	$73.44_{\pm 0.73}$
IF-MMIN (Zuo et al., 2022)	ResNet-101	image	$72.58_{\pm 0.48}$
Class-name prompt and matching baselines			
CoOP (Zhou et al., 2022b)	CLIP	image + class text	$62.34_{\pm 0.64}$
CoCoOp (Zhou et al., 2022a)	CLIP	image + class text	$71.53_{\pm 1.01}$
CLIP label match (Radford et al., 2021)	CLIP	image + class text	$12.78_{\pm 0.00}$
Detection-based image baseline			
YOLO11n detection → image-level classification	YOLO11n	image	$84.52_{\pm 0.71}$
Retained generated-view model			
SEAR	ResNet-101 + Qwen2.5-VL	image + generated text	$\mathbf{88.15_{\pm 0.57}}$

tage depends on the description budget: human-authored text performs better with one, three, or ten descriptions. Additional descriptions can supply label-relevant attributes but also introduce redundant or inconsistent content, affecting the balance between structural benefit and estimation error in Eq. (12). The description budget should therefore be selected for its predictive value rather than simply maximized.

D.4 Semantic Recoding Example

Figure 1 illustrates the qualitative mechanism behind semantic recoding. Although the defect instances differ in location, texture, and local appearance, the generated descriptions repeatedly map them to stable semantic attributes, giving the classifier a more standardized textual coordinate for visually heterogeneous evidence.

D.5 Synthetic Check Details

The synthetic experiment is a qualitative check of the structural-gain-versus-estimation-cost trade-off in Eq. (12), rather than an exact simulation of the nonparametric framework in Section 4. We set $\beta_u = \beta_t = 1$, $\beta_a = (1, 0, 0)^\top$, $\beta_s = 0$, $\sigma_y = 0.5$, image noise $\sigma_t = 0.7$, and a generated view with $k = 3$ regressors. Thus one generated coordinate is label-aligned and the remaining coordinates contribute estimation cost. The residual variance σ_r^2 is computed from the best image-only linear predictor and includes the residual effect of noisy image observation. For each pair of generator noise $\sigma_{\mathcal{G}}$ and sample size n, we fit image-only and fused linear predictors by OLS and average the test MSE difference over 200 training draws, using a large independent test set of size 2×10^5. The purpose is to check the sign and monotonic trends suggested by the sufficient condition $\Delta > r_f(n)$; it is not a numerical verification of a closed-form boundary.

D.6 Detailed Steel Defect Baselines

Table 5 extends Table 3 with the backbone and inference input of each method. Training-time language-use methods transfer generated-text supervision into an image-only classifier; SMIL and IF-MMIN also use image-only inference through their missing-modality formulations. Class-name prompt and matching methods operate on category-level text rather than instance-level descriptions.

Table 6: Additional image-only and rule-based text controls. Results are F1-macro (%).

Dataset	CLIP Image	ViT Image	Rule Text
Steel Defect	70.78±1.37	83.76±0.70	31.28±0.00
EuroSAT	92.92±0.14	94.29±0.15	11.06±0.00
Blood Cell	97.69±0.20	98.38±0.26	27.11±0.00
Bone Fracture	91.20±0.01	87.50±0.04	51.39±0.00
Pneumonia	93.72±0.35	93.23±0.48	23.64±0.00
Beans	91.81±0.25	93.58±0.53	37.34±0.00
Oxford Pets	95.59±0.01	94.34±0.80	62.88±0.00
Cats vs Dogs	99.25±0.25	99.51±0.40	100.00±0.00

Table 7: Additional comparison with test-time multimodal fusion methods adapted to image and generated-text features. Results are F1-macro (%).

Dataset	RM-CBP	MCB	MFN	MVAE	SEAR
Steel Defect	55.93±0.97	57.13±0.62	66.42±1.79	71.38±0.60	**88.15±0.57**
EuroSAT	85.82±0.20	85.46±0.09	92.25±0.38	92.93±0.10	**96.99±0.34**
Blood Cell	94.26±0.16	92.90±0.29	96.94±0.13	98.05±0.09	**99.07±0.34**
Bone Fracture	84.51±0.35	83.14±0.13	87.15±2.82	91.36±0.65	**92.08±0.76**
Pneumonia	89.73±0.24	89.40±0.21	91.47±1.00	92.02±1.24	**95.66±0.23**
Beans	88.43±0.45	88.57±0.28	91.14±0.53	93.30±1.19	**96.29±0.83**
Oxford Pets	93.41±0.13	93.36±0.79	93.81±0.89	95.65±0.40	**95.72±0.62**
Cats vs Dogs	98.50±0.42	98.84±0.61	99.15±0.48	98.97±0.85	**99.33±0.47**

D.7 Additional Image and Text Controls

Table 6 reports supplementary controls omitted from the main results table for compactness. CLIP IMAGE and VIT IMAGE replace the visual feature source while keeping image-only inference, and therefore serve as image-only references rather than SEAR variants. RULE TEXT predicts directly from generated descriptions using fixed keyword rules for class names or diagnostic phrases. These controls help separate backbone choice and explicit text parsing from the SEAR fusion mechanism; the perfect Cats vs Dogs rule score reflects that the generated descriptions almost always name the animal category.

D.8 Additional Multimodal-Fusion Baselines

Table 7 reports test-time multimodal fusion methods originally developed for settings with naturally observed auxiliary modalities, including RM-CBP (Gao et al., 2016), MCB (Fukui et al., 2016), MFN (Zadeh et al., 2018), and MVAE (Wu & Goodman, 2018). We adapt them to the same image features and default 7B generated-text features used by SEAR on every dataset, so the comparison isolates fusion design under the generated-view setting rather than differences in available information. As discussed in Section 4, independently observed and generated text can differ in the information they provide, motivating a separate comparison of these fusion methods. Because these baselines were designed for naturally observed auxiliary modalities, the table should be read as a diagnostic compatibility check rather than a claim about their original use cases.

E Prompt Templates

Table 8 lists the instruction ρ_{instr} used on every dataset. Examples are added only on Steel Defect ($K = 7$); other main-table datasets use the instruction alone (Section 3.2). The final row records the bird-attribute prompt used for the CUB-20 diagnostic in Section D.3.

F Cost-Aware Pilot Assessment

Generated descriptions have dataset-dependent utility and generator cost, so we include a cost-aware pilot assessment as an engineering diagnostic for deciding before full training whether to use

Table 8: Instruction ρ_{instr} of the generation prompt. Examples are added only on Steel Defect ($K = 7$); other main-table datasets use the instruction alone (Section 3.2).

Dataset	Prompt ρ
Steel Defect	System: "You are an expert in steel surface inspection and defect analysis. Given an image of a steel surface, describe it in one concise and professional paragraph in English. The description should mention the surface tone, contrast, visible texture, and any visible defect characteristics, including appearance, formation cause, and general assessment of the surface condition. Below are several examples:" Current: "Now analyze the following image and produce a description in the same style:"
EuroSAT	"This is a remote sensing satellite image from a land use classification dataset. Please describe the main land cover types, man-made structures, vegetation, water, and any notable patterns or textures you can observe from the top-down view."
Pneumonia	"This is a chest X-ray image from a medical imaging dataset. Please describe the appearance of the lungs, any visible opacities, consolidations, or abnormal patterns, as well as the overall condition of the chest area. Do not make a medical diagnosis; only describe what you see in the image."
Beans	"This is an image of bean leaves. Please describe the leaf condition, including any spots, rust, or other visible issues."
Cats vs Dogs	"This is an image of a cat or a dog. Please describe the animal, its appearance, and pose."
Blood Cell	"This is an image of blood cells from a microscopy dataset. Please describe the blood cell appearance, morphology, and any visible abnormalities or features that might be relevant for classification."
Bone Fracture	"This is a medical X-ray image of a bone. Please describe the bone structure, any visible fractures, cracks, or abnormalities, and the overall condition of the bone."
Oxford Pets	"This is an image of a cat or dog from a pet dataset. Please describe the pet's breed if possible, and its appearance."
CUB-20 diagnostic	"This is an image of a bird from a fine-grained bird classification dataset. Please describe species-relevant visual attributes such as plumage color, beak shape, wing and tail appearance, body shape, pose, and visible markings."

Table 9: Cost-aware pilot assessment over 50 random pilot splits. Entries report the fraction of pilot runs selecting each candidate.

Dataset	Image-only	3B text	7B text
Steel Defect	28%	20%	52%
EuroSAT	66%	26%	8%
Blood Cell	56%	14%	30%
Bone Fracture	36%	38%	26%
Pneumonia	40%	30%	30%
Beans	34%	56%	10%
Oxford Pets	30%	6%	64%
Cats vs Dogs	8%	22%	70%

generated text and which generator scale to try. The procedure is a deployment rule rather than theoretical evidence for the sufficient condition in Eq. (12): it operationalizes the ω-side choice of generator scale and is not an estimate of a theoretical generator-noise threshold. For each dataset, we draw 50 random pilot splits, train an IMAGE-ONLY candidate and two generated-text candidates using 3B or 7B text, and select the candidate with the highest pilot-validation F1-macro. The reported main experiments do not use this pilot selector; they report the fixed configurations described in Section 5.1, with explicit 3B diagnostics only where stated.

Concentrated selections indicate a clear pilot-level deployment choice for Cats vs Dogs and Oxford Pets, which mostly select 7B text, and Beans, which mostly selects the lower-cost 3B text. Steel Defect favors 7B text but shows a less concentrated profile, with image-only also selected in a substantial fraction of pilot splits. Dispersed selections on Pneumonia and Bone Fracture indicate

small or unstable pilot gains, while EuroSAT and Blood Cell lean toward image-only selection. The profile should be read as a pre-training decision aid rather than as evidence for the closed-form threshold or as an explanation of the reported full-scale results.

Algorithm 1 Cost-aware pilot assessment

Require: Training split, candidate set $\mathcal{C} = \{\text{IMAGE-ONLY}, \text{3B TEXT}, \text{7B TEXT}\}$, pilot budget b, number of splits S
Ensure: Selection profile over candidates
1: **for** each pilot split $s = 1, \ldots, S$ **do**
2: Draw up to b samples per class and split them into pilot train and pilot validation sets.
3: **for** each candidate $c \in \mathcal{C}$ **do**
4: Train c on the pilot train set under the same evaluation protocol.
5: Evaluate F1-macro $F_s(c)$ on the pilot validation set.
6: **end for**
7: Record $c_s^\star \leftarrow \arg\max_{c \in \mathcal{C}} F_s(c)$.
8: **end for**
9: Report $\frac{1}{S} \sum_{s=1}^{S} \mathbb{I}[c_s^\star = c]$ for each $c \in \mathcal{C}$.

G Optimal Hyperparameters

Table 10 reports the optimal hyperparameter configuration for the full model (SEAR) on each dataset, selected via Optuna Akiba et al. (2019). The threshold η is a single shared validation-tuned threshold for the multi-label Steel Defect task; it is not used for single-label datasets.

Table 10: Optimal hyperparameters for the full model (SEAR) on each dataset. Threshold η is used only for the multi-label Steel Defect task.

	Steel	EuroSAT	Blood Cell	Bone Frac.	Pneumonia	Beans	Oxford Pets	Cats/Dogs
Learning rate	1.498e-04	1.23e-04	3.47e-04	3.39e-04	1.62e-04	1.36e-03	7.39e-05	1.03e-04
Threshold η	0.6409	–	–	–	–	–	–	–
Dropout (layer 1)	0.3503	0.2664	0.4275	0.2244	0.3422	0.3317	0.4997	0.4709
Dropout (layer 2)	0.2277	0.4225	0.4125	0.4088	0.1709	0.4812	0.3354	0.4350
Text projection dropout	0.4236	0.3184	0.1678	0.3465	0.4625	0.4949	0.3852	0.4698
λ_{align}	0.1081	0.1060	0.8305	0.3293	0.1855	0.7369	0.5833	0.4831
τ	0.0339	0.1257	0.0604	0.0354	0.0897	0.0174	0.1655	0.0582
λ_{rec}	0.3384	1.6984	1.1261	1.8180	0.5535	0.2065	1.4313	0.8748