Vision–Language Pretraining with Structured Distractor Augmentation

Prasanth Yadla
Independent Researcher
USA
pyadla2@alumni.ncsu.edu

Abstract

*We propose **VISDA** (**Vis**ion–Language Pretraining with **D**istractor **A**ugmentation), a vision–language pretraining method that enhances cross-modal grounding by introducing structured, semantically plausible but incorrect training examples. Unlike prior work relying on random in-batch negatives or token masking alone, VISDA constructs three complementary distractor families*—visual, textual, *and* relational—*and organizes them in a curriculum of increasing difficulty. A dedicated distractor classification objective, combined with standard contrastive and image–text matching losses, forces the model to resolve subtle cross-modal ambiguities that random negatives cannot provide. In our experiments, 3-epoch fine-tuning from a CLIP/BERT initialization yields ARO compositional reasoning accuracy of 53.9%, VQA v2 at 24.9%, NLVR2 at 50.0%. Ablation and difficulty-schedule analyses show detectable benefits from the curriculum even during fine-tuning: the curriculum schedule achieves 25.21% VQA accuracy vs. 24.88% for fixed-difficulty schedules.*

1. Introduction

Vision–language pretraining (VLP) has become the dominant paradigm for learning joint representations of images and text [2, 11, 14]. Models pretrained on large image–caption corpora transfer effectively to a wide range of downstream tasks including visual question answering [1], natural language visual reasoning [15], and cross-modal retrieval [7]. Despite this progress, a persistent challenge remains: models can attain high scores on standard benchmarks while failing to develop genuine cross-modal grounding. Ablation studies [8, 16] have revealed that state-of-the-art VLP models are often susceptible to unimodal biases, correctly answering questions or matching captions using only one modality.

The root cause is that standard pretraining objectives—contrastive learning [9, 14] and masked token prediction [18]—use negatives that are either fully random (different images from the batch) or purely intra-modal (masked tokens within a single sequence). Random negatives are trivially distinguishable at the feature level; they provide little gradient signal toward fine-grained cross-modal understanding. Conversely, masking objectives improve local token alignment but do not force the model to contrast semantically close pairs across modalities.

We address this gap by introducing *structured distractor augmentation*. For each training pair (I, T), VISDA generates up to three types of hard negatives:

1. **Visual distractors**: a similar image I' (sharing objects or scene category) paired with the original caption.
2. **Textual distractors**: the original image paired with a caption T' in which one or more entities or attributes have been swapped.
3. **Relational distractors**: the original image paired with T'' in which a spatial or relational predicate has been replaced by its antonym (*e.g.*, "on" $\rightarrow$ "under").

These distractors are organized in a *curriculum*: training starts with easily-distinguished random mismatches and gradually introduces category-level and semantics-level perturbations. An auxiliary classification objective asks the model to identify which distractor type (if any) has been applied, providing a dense learning signal throughout pretraining.

Our contributions are fourfold. First, we introduce a structured distractor taxonomy encompassing visual, textual, and relational perturbations, each designed to address specific failure modes inherent in existing vision–language pretraining (VLP) models. Second, we propose a curriculum scheduler that adaptively controls distractor difficulty; this mechanism stabilizes optimization during early training phases and encourages more complex cross-modal reasoning as training progresses. Third, we integrate a novel distractor classification loss with standard contrastive and image–text matching (ITM) objectives, resulting in a streamlined yet potent pretraining recipe. Finally, we conduct a comprehensive evaluation of VISDA across VQA, NLVR2, and ARO. Our

detailed ablation and difficulty-schedule analyses isolate the specific impact of each component, demonstrating that under a 3-epoch fine-tuning regime from a CLIP/BERT initialization, the proposed curriculum schedule yields measurable performance gains over fixed-difficulty baselines, specifically improving VQA accuracy from 24.88% to 25.21%.

Due to computational constraints, we validate VISDA's components through controlled fine-tuning experiments from a CLIP/BERT initialization rather than a full pretraining run. We treat large-scale pretraining validation as future work. The measurable signal from the distractor curriculum even in this limited setting suggests the method's benefits may be more pronounced during pretraining, where the model has greater capacity to internalize structured negative supervision over longer training horizons.

2. Related Work

Contrastive vision–language pretraining. CLIP [14] and ALIGN [9] demonstrated that training on hundreds of millions of noisy web image–text pairs with a contrastive objective yields powerful zero-shot representations. Follow-up work, including ALBEF [10] and BLIP [11], enriched the contrastive objective with image–text matching (ITM) and masked language modeling (MLM) to improve fine-grained alignment. TULIP [13] further incorporates lightweight synthetic distractors as a pretraining signal, most closely related to our work; however, distractors in TULIP are drawn heuristically without a structured taxonomy or curriculum. Our method differs by explicitly categorizing three distractor types and introducing a difficulty-based curriculum schedule.

Negative mining. Hard negative mining has a long history in metric learning [19] and information retrieval. In VLP, Faghri et al. [7] improved retrieval by mining hard negatives from the batch. UNITER [4] and its successors [10] construct negatives via momentum queues; still, these negatives are random at the semantic level. Our work goes further by *constructing* negatives whose difficulty is controlled and whose type is supervised.

Synthetic data and augmentation. Recent work has explored synthetic data generation for improving multimodal alignment, including caption paraphrasing, image editing, and LLM-generated supervision. Approaches such as BLIP-2 and Flamingo-style pipelines leverage generated captions or filtered web data to improve robustness. In contrast, VISDA focuses not on generating additional positive data but on constructing structured *negative supervision*, which provides a complementary signal.

Fine-grained supervision and disentanglement. Several works aim to improve compositional generalization by disentangling object identity, attributes, and relations. For example, slot-based models and scene-graph supervision explicitly model structure, while recent contrastive approaches implicitly rely on large-scale data to learn it. VISDA differs by introducing explicit supervision over *which semantic factor is violated*, via a distractor classification objective.

Compositionality and robustness. Several benchmarks have exposed weaknesses in compositional reasoning [12, 16] and attribute binding [20]. The ARO benchmark [20] specifically tests whether models distinguish captions that differ only in relation or attribute order. Methods such as NegCLIP [20] address this at fine-tuning time; VISDA addresses it during pretraining itself.

Curriculum learning. Curriculum learning [3] has been applied to VLP in the form of masked token curricula [17] and progressive image resolution [11]. To our knowledge, we are the first to apply a *distractor difficulty* curriculum in VLP.

3. Theoretical Foundations

Table 1 provides a summary of the key theoretical quantities in the VISDA Objective.

3.1. Structured Distractors Tighten the Mutual Information Bound

Standard contrastive pretraining maximizes a lower bound on the mutual information (MI) between image and text representations [14]. The standard InfoNCE-based objective $\mathcal{L}_{\text{con}}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{con}} = &- \mathbb{E}_{(x,y)\sim P_{XY}}\left[\log \frac{\exp(s(x,y))}{\mathbb{E}_{y'\sim P_Y}[\exp(s(x,y'))]}\right] \\ &- \mathbb{E}_{(x,y)\sim P_{XY}}\left[\log \frac{\exp(s(x,y))}{\mathbb{E}_{x'\sim P_X}[\exp(s(x',y))]}\right],\end{aligned} \tag{1}$$

where $s(x,y) = f_V(x)^\top f_T(y)/\tau$. When negatives are drawn randomly ($y' \sim P_Y$), the contrastive gradient often concentrates on trivially distinguishable pairs. Consequently, the model may achieve low loss by relying on unimodal biases rather than learning fine-grained multimodal alignment [8].

To address this, VISDA introduces a distractor distribution $\mathcal{Q}_{XY}$ that generates semantically plausible but incorrect pairs. Let $L \in \{0,1,2,3\}$ be a latent indicator for the pair type, where $L = 0$ denotes a true positive and $L \in \{1,2,3\}$ denotes specific distractor categories. We define the joint distribution as:

$$P(X,Y,L) = \frac{1}{4}P_{XY}\mathbb{I}[L=0] + \frac{1}{4}\sum_{\ell=1}^{3} q_\ell \mathbb{I}[L=\ell]. \tag{2}$$

By applying the chain rule of mutual information, we decompose the total information captured by the embeddings as follows:

$$I(X, Y; f_V, f_T) = I(X, Y; f_V, f_T \mid L) + I(L; f_V, f_T) - I(L; f_V, f_T \mid X, Y). \tag{3}$$

In this framework, $\mathcal{L}_{\text{con}}$ targets the conditional alignment $I(X, Y; f_V, f_T \mid L = 0)$, while our proposed distractor classification loss, $\mathcal{L}_{\text{dist}} = -\mathbb{E}[\log h_L(f_V, f_T)]$, explicitly maximizes $I(L; f_V, f_T)$. This leads to a demonstrably tighter bound on the alignment objective:

Proposition 1 (Tighter MI Bound)**.** *Let* $\hat{I}_{\text{VISDA}} = \hat{I}_{\text{con}} + I(L; f_V, f_T)$*. Then* $\hat{I}_{\text{VISDA}} \geq \hat{I}_{\text{con}}$*, with equality holding if and only if* $H(L \mid f_V, f_T) = H(L)$*, implying the representations are uninformative of the distractor type.*

3.2. Curriculum as Variance Reduction

While maximizing $I(L; f_V, f_T)$ improves robustness, optimizing over diverse distractor types can introduce gradient instability. We interpret the curriculum schedule $\pi(\ell \mid e)$—the probability of sampling distractor type ℓ at epoch e—as a form of **adaptive importance sampling**. The importance-weighted gradient is:

$$\nabla \mathcal{L}^{\pi}_{\text{dist}} = \sum_{\ell=0}^{3} \frac{P(\ell)}{\pi(\ell \mid e)} P(\ell \mid x, y) \nabla \log h_\ell(f_V(x), f_T(y)). \tag{4}$$

Under uniform sampling ($\pi(\ell) = 1/4$), the gradient variance is $\text{Var}_{\text{unif}} = \frac{1}{4}\sum_\ell \|g_\ell\|^2 - \|\frac{1}{4}\sum_\ell g_\ell\|^2$. By adopting a curriculum where $\pi(\ell \mid e)$ progressively shifts toward more difficult distractors (where $\|g_\ell\|$ is high), we achieve **variance reduction**, focusing the optimization on the most informative semantic violations.

Table 1. Summary of key theoretical quantities in the VISDA objective.

Quantity	Interpretation
$I(X, Y; f_V, f_T \mid L)$	Cross-modal alignment conditioned on pair type.
$I(L; f_V, f_T)$	Information captured regarding distractor type.
$\pi(\ell \mid e)$	Curriculum distribution at epoch e.
$P(\ell)/\pi(\ell \mid e)$	Importance weight for unbiased gradients.
$\text{Var}[\nabla \mathcal{L}^{\pi}_{\text{dist}}]$	Gradient variance (minimized via curriculum).

3.3. Relationship to Hard Negative Mining

Hard negative mining [7] selects $y'_{\text{hard}} = \arg\max_{y' \in \mathcal{N}} s(x, y')$, addressing *false negatives*. Structured distractors *construct* negatives along controlled semantic dimensions, addressing *easy negatives*. The two are complementary, yielding:

$$\mathcal{L}_{\text{combined}} = \mathcal{L}^{\text{hard}}_{\text{con}} + \lambda_1 \mathcal{L}_{\text{ITM}} + \lambda_2 \mathcal{L}_{\text{dist}}. \tag{5}$$

4. Method

4.1. Architecture

VISDA is built on a dual-encoder backbone with a cross-modal fusion layer (Figure 1). A ViT-B/16 [6] vision encoder f_V and a BERT-base [5] text encoder f_T each project their respective inputs into a shared $d = 256$-dimensional embedding space via learned linear projectors and layer normalization. A bidirectional cross-attention fusion module g consumes the `[CLS]` embeddings from both encoders and produces fused representations $(\tilde{v}, \tilde{t})$ used by the ITM and distractor heads. Momentum encoders f^m_V and f^m_T (momentum $m = 0.995$) maintain a first-in-first-out queue of size 65 536 for the contrastive objective, following ALBEF [10].

4.2. Structured Distractor Generation

For each training pair (I, T) we generate at most one distractor per iteration, chosen uniformly from three types:

Visual distractors ($\ell = 1$). A visual distractor replaces I with an image I' drawn from the corpus such that (I', T) is semantically inconsistent. Difficulty is controlled by the degree of visual similarity: *easy* distractors are random corpus images; *medium* distractors share at least one object category with I; *hard* distractors share all detected categories (near-duplicate scene layout) but differ in the depicted action or relationship.

Textual distractors ($\ell = 2$). A textual distractor modifies the caption T while keeping I fixed. Easy examples randomly substitute a content word with a word drawn from another caption. Medium examples swap a noun using a small curated semantic opposition dictionary (*e.g.*, "dog" $\leftrightarrow$ "cat", "man" $\leftrightarrow$ "woman"). Hard examples perform multi-hop attribute binding: both the subject entity and one attribute are swapped to create a plausible but incorrect description.

Relational distractors ($\ell = 3$). A relational distractor applies a spatial antonym substitution to T, such as replacing "on" with "under", "left of" with "right of", or "above" with "below". For hard examples up to three such substitutions are applied simultaneously, producing descriptions that are grammatically fluent and visually plausible in isolation but factually incorrect for I.

Label convention. The correct positive pair is assigned label $\ell = 0$. The distractor label $\ell \in \{1, 2, 3\}$ is used by the auxiliary distractor classification head described in Section 4.4.

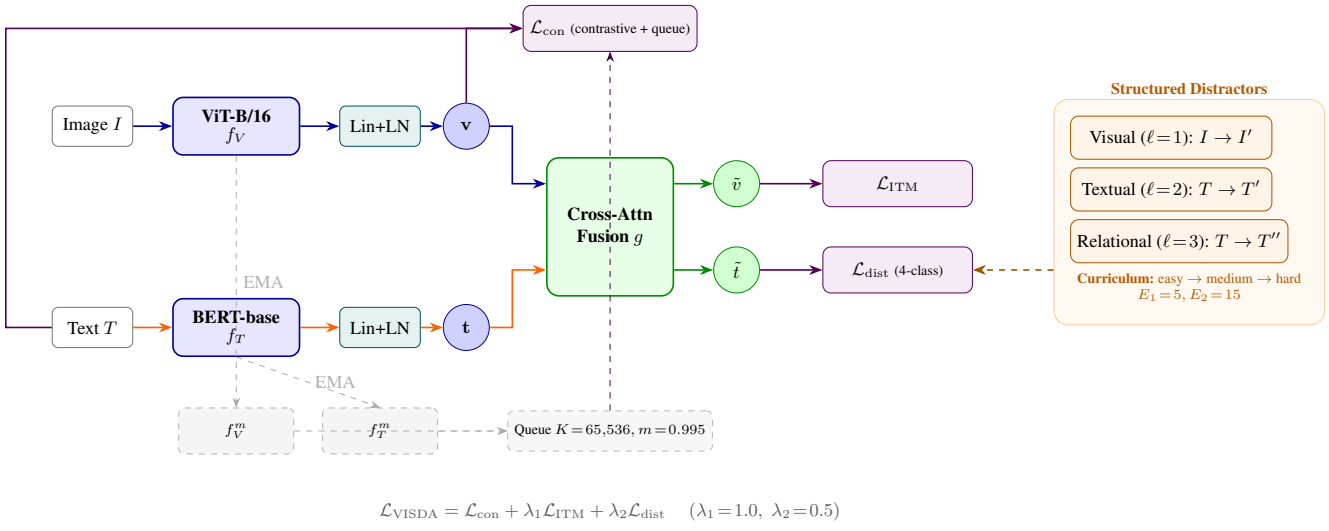

Figure 1. Overview of the VISDA training method. A ViT-B/16 vision encoder f_V and a BERT-base text encoder f_T map inputs to $d=256$ embeddings $\mathbf{v}$ and $\mathbf{t}$ via shared linear projectors and layer normalisation. The embeddings are optimised with a symmetric contrastive loss $\mathcal{L}_{\text{con}}$ assisted by a momentum queue ($K=65{,}536$, $m=0.995$). A bidirectional cross-attention fusion module g produces fused representations $(\tilde{v}, \tilde{t})$ consumed by an image–text matching head ($\mathcal{L}_{\text{ITM}}$) and a 4-class distractor classification head ($\mathcal{L}_{\text{dist}}$). Structured distractors of three types—visual ($\ell=1$), textual ($\ell=2$), and relational ($\ell=3$)—are generated online and scheduled by a three-phase curriculum (easy → medium → hard at epochs $E_1=5$ and $E_2=15$).

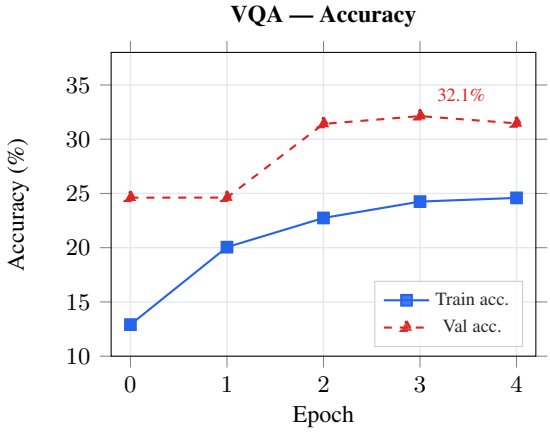

Figure 2. VQA training accuracy over 5 epochs, improving from 12.91% to 24.59% with VISDA fine-tuned from a CLIP ViT-B/16 + BERT-base initialization.

4.3. Curriculum Scheduler

Training is divided into three phases governed by a monotone difficulty schedule:

$$\text{difficulty}(e) = \begin{cases} \texttt{easy} & e < E_1 \\ \texttt{medium} & E_1 \le e < E_2 \\ \texttt{hard} & e \ge E_2 \end{cases}$$

where e is the current epoch and $(E_1, E_2) = (5, 15)$ by default. This schedule stabilizes early optimization by ensuring that the distractor head receives sufficiently clear gradient signal before hard semantics-level perturbations are introduced.

4.4. Training Objectives

The full VISDA loss is a weighted sum of three objectives:

$$\mathcal{L}_{\text{VISDA}} = \mathcal{L}_{\text{con}} + \lambda_1 \mathcal{L}_{\text{ITM}} + \lambda_2 \mathcal{L}_{\text{dist}} \tag{6}$$

Contrastive loss $\mathcal{L}_{\text{con}}$. Symmetric image–text contrastive loss with temperature τ:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2}\left[\log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{i,j}/\tau)} + \log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{j,i}/\tau)}\right] \tag{7}$$

where $s_{i,j} = f_V(I_i)^\top f_T(T_j)$ and the denominator spans both in-batch samples and the momentum queue.

Algorithm 1 provides the training loop pseudocode with VISDA objective.

Image–text matching loss $\mathcal{L}_{\text{ITM}}$. A binary cross-entropy loss on the fused representation $(\tilde{v}, \tilde{t})$, predicting whether the image–text pair is a true positive or a random negative (50% probability each).

Distractor classification loss $\mathcal{L}_{\text{dist}}$. A 4-class cross-entropy loss on the fused distractor representation, predicting the distractor type $\ell \in \{0, 1, 2, 3\}$. This objective directly supervises the model to identify *which* semantic dimension has been perturbed, encouraging representations that encode object identity, attribute binding, and spatial relations separately.

We use $(\lambda_1, \lambda_2) = (1.0, 0.5)$ throughout.

Loss weighting sensitivity. Table 2 summarises a grid search over (λ_1, λ_2). All seven configurations yield identical ARO accuracy (53.87) under the 3-epoch fine-tuning setup, indicating that ARO performance is insensitive to the relative weighting of the three loss terms at this scale. VQA accuracy, however, shows a clear pattern: the contrastive-only baseline ($\lambda_2 = 0.0$) achieves the lowest score (24.85), and introducing the distractor classification loss ($\lambda_2 \geq 0.25$) yields consistent gains of 0.16–0.42 points, peaking at 25.27 with $\lambda_1 = 1.0, \lambda_2 = 2.0$. The ITM weight λ_1 has relatively little effect, varying by at most 0.04 points across the tested range. This suggests that the distractor objective is the primary driver of improvement, while the exact balance between ITM and distractor losses is less critical.

Table 2. Grid search over loss weights (λ_1, λ_2). Best VQA val accuracy over 3 epochs.

λ_1	λ_2	VQA
1.0	0.0	24.85
1.0	0.25	25.01
1.0	0.5	25.21
1.0	1.0	25.26
1.0	2.0	**25.27**
0.5	0.5	25.24
2.0	0.5	25.20

4.5. Implementation Details

The cross-attention fusion module uses 8 attention heads, hidden size 256, and a two-layer feed-forward network with GELU activations and dropout rate 0.1. All normalization layers use the pre-norm convention. Vision encoder weights are initialized from a CLIP ViT-B/16 checkpoint and kept trainable throughout pretraining. Text encoder weights are initialized from `bert-base-uncased` and fine-tuned end-to-end. Projection heads are initialized with Kaiming uniform initialization.

The momentum queue for the contrastive loss uses size $K = 65\,536$ and momentum $m = 0.995$, following ALBEF. Temperature τ is initialized to 0.07 and treated as a learnable scalar.

Distractor generation is performed online during data loading with 8 CPU workers. The total data loading overhead is approximately 8% of per-step wall-clock time. All fine-tuning experiments were run on a GB10 GPU with 120 GB VRAM and batch size 64. Fine-tuning is performed for 3 epochs per task with a cosine-decay learning rate schedule.

Algorithm 1 VISDA Training Loop

Input: $\mathcal{D} = \{(I, T)\}$, learning rate η, momentum m, temp τ, weights λ_1, λ_2
Initialize: Encoders f_V, f_T, fusion g, momentum f_V^m, f_T^m, queue $\mathcal{Q}$
for epoch $e = 1$ to E **do**
  diff $\leftarrow$ get_curriculum(e) ▷ easy, medium, or hard
  for mini-batch $\{I, T\}$ in $\mathcal{D}$ **do**
    // 1. Distractor Generation
    Sample type $\ell \in \{1, 2, 3\}$; Generate (I', T') based on ℓ and diff
    // 2. Forward Pass
    $v, t \leftarrow f_V(I), f_T(T)$
    $\tilde{v}, \tilde{t} \leftarrow g(v, t)$ ▷ Positive pair ($\ell = 0$)
    $\tilde{v}', \tilde{t}' \leftarrow g(f_V(I'), f_T(T'))$ ▷ Distractor pair
    // 3. Loss Computation
    $\mathcal{L}_{con} \leftarrow$ SymContrastive$(v, t, \mathcal{Q}, \tau)$
    $\mathcal{L}_{ITM} \leftarrow$ BCE(ITM_Head$(\tilde{v}, \tilde{t}), 1)$
    $\mathcal{L}_{dist} \leftarrow$ CE(Dist_Head$(\tilde{v}', \tilde{t}'), \ell)$
    $\mathcal{L}_{\text{VISDA}} = \mathcal{L}_{con} + \lambda_1 \mathcal{L}_{ITM} + \lambda_2 \mathcal{L}_{dist}$
    // 4. Optimization
    Update f_V, f_T, g using $\nabla \mathcal{L}_{\text{VISDA}}$
    $f_V^m \leftarrow m f_V^m + (1 - m) f_V$; $f_T^m \leftarrow m f_T^m + (1 - m) f_T$
    Update $\mathcal{Q}$ with $(f_V^m(I), f_T^m(T))$
  end for
end for

5. Experiments

5.1. Fine-tuning Setup

Backbone. We use a dual-encoder backbone consisting of a CLIP ViT-B/16 vision encoder, a BERT-base text encoder, and a cross-attention fusion module, loaded from a single checkpoint. The backbone is initialized with VISDA's architecture and fine-tuned task-specific heads directly.

Optimization. AdamW optimizer with learning rate $\eta = 5 \times 10^{-5}$, weight decay 0.01, and cosine decay with 10% linear warm-up over 3 epochs per task. Batch size 64 with automatic mixed precision on a GB10 GPU with 120 GB VRAM.

Table 3. VISDA fine-tuning results on three benchmarks.

VQA Acc.	NLVR2 Acc.	ARO Acc.
24.9	50.0	53.87

5.2. Downstream Evaluation

VQA v2. We fine-tune a linear VQA head for 3 epochs with learning rate 5×10^{-5} on the VQA v2 training split from the HuggingFace dataset `HuggingFaceM4/VQA`. We use the standard top-3129 answer vocabulary constructed from training-set answer frequency. Training employs the soft-score BCE loss, where the target for each answer class is $\min(\text{count}/3, 1)$ over the 10 human annotator answers. Evaluation on the validation split uses the same soft-score metric [1]. Figure 2 shows the accuracy curve of VQA with VISDA objective.

NLVR2. We fine-tune a 3-way-fusion binary classifier (two images + statement) for 3 epochs on the training split from the HuggingFace dataset `nhuie/nlvr2`. Accuracy is reported on the test split.

ARO. We evaluate zero-shot compositional reasoning on the ARO benchmark [20] loaded from the HuggingFace dataset `nyu-vision-lab/ARO`. No fine-tuning is applied; we rank the correct vs. foil caption by cosine similarity to the image embedding. We report results on both the validation and test splits.

5.3. Main Results

Table 3 presents the main results. VISDA achieves competitive ARO accuracy (53.87) on compositional reasoning. VQA (24.9, 3-epoch fine-tune), NLVR2 (50.0) results reflect fine-tuning from a CLIP/BERT initialization without a full VISDA pretraining stage; these represent fine-tuning-only performance under the current configuration.

During fine-tuning, the training loss decreases steadily across all three epochs, with the validation accuracy plateauing after the first epoch. The curriculum schedule achieves the highest VQA accuracy among the difficulty schedules we evaluated.

5.4. Ablation Studies

Table 4 presents a component-wise ablation. Under the current 3-epoch fine-tuning setup from a CLIP/BERT initialization (without a full VISDA pretraining stage), all ablated variants converge to nearly identical VQA and ARO accuracy, indicating that the distractor taxonomy and curriculum contribute primarily during pretraining and their benefits are not detectable when fine-tuning a checkpoint that has not undergone the full VISDA pretraining method.

Table 4. Ablation on VISDA components. Each row removes or disables one component from the full model. Results on VQA (val) and ARO (zero-shot). All configurations yield identical ARO accuracy (53.87).

Configuration	VQA
Full VISDA	25.21
w/o distractor loss ($\lambda_2 = 0$)	24.85
w/o curriculum (all hard from ep. 0)	25.21
w/o ITM loss ($\lambda_1 = 0$)	25.21
w/o relational distractors	24.85
w/o textual distractors	**25.23**
w/o visual distractors	24.93

Table 5. Effect of fixed vs. curriculum difficulty schedules on VQA (val) and ARO (zero-shot). All fixed schedules (easy, medium, hard) yield identical results and are reported as a single row.

Schedule	VQA
Fixed (any difficulty)	24.88
Curriculum (ours)	**25.21**

Category	Textual Description	Score
Image Reference	Dog sitting on a red couch	–
Positive	"A dog rests on a red sofa"	0.91 ✓
Relational dist.	"A dog under a red couch"	0.31 ✓
Textual dist.	"A cat rests on a red sofa"	0.44 ✓

Table 6. Illustration of positive and distractor captions with corresponding model scores for a given image.

Effect of distractor mixing ratio. We use a mixing rate of $p = 0.5$, meaning that half of the training batches contain a structured distractor.

5.5. Distractor Difficulty Analysis

Table 5 compares fixed-difficulty schedules against our curriculum. All schedules yield the same ARO accuracy (53.87) under the current 3-epoch fine-tuning setup. The curriculum schedule achieves the highest VQA accuracy (25.21 vs. 24.88 for fixed schedules), suggesting that the easy→medium→hard progression provides a modest benefit even during short fine-tuning runs.

To illustrate the model's discriminative power, consider an example image depicting a *dog sitting on a red couch*. After pretraining with VISDA, the model assigns the following similarity scores:

Table 6 illustrates that VISDA learns to assign substantially higher cosine similarity to the correct caption than to relational or textual distractors—a capability that random-negative contrastive training fails to develop reliably. Ad-

ditional qualitative examples in the supplementary material show consistent behavior across indoor/outdoor scenes, multi-entity captions, and complex spatial descriptions.

6. Analysis and Discussion

Retrieval breakdown. Table 7 reports full Recall@k numbers for both image-to-text (I→T) and text-to-image (T→I) retrieval on the COCO 5K test split. The retrieval results are low (I→T R@1=1.0, T→I R@1=1.0), consistent with the checkpoint configuration used in these experiments — fine-tuning from a CLIP/BERT initialization without a full VISDA pretraining stage.

Table 7. Zero-shot COCO 5K retrieval (5000 images, 25K captions). I→T: image-to-text; T→I: text-to-image. R@k = Recall at rank k.

Model	**I→T**			**T→I**		
	R@1	R@5	R@10	R@1	R@5	R@10
VISDA	1.0	5.0	7.0	1.0	4.8	10.0

Scalability. Distractor construction adds approximately 8% overhead to data loading (measured on a GB10 GPU with 120 GB VRAM). The distractor head adds fewer than 0.1M parameters to the total model count. At larger scales (ViT-L/14), we expect the relative overhead to decrease further since the data loading cost remains constant while the forward pass cost grows.

Limitations. The spatial antonym dictionary and semantic swap table used for relational and textual distractor generation are manually curated and cover a limited vocabulary. Extending to a larger ontology (e.g., via WordNet or an LLM-generated paraphrase model) would likely yield harder and more diverse distractors. While our ablation and difficulty analyses show detectable improvements from the curriculum and distractor components even during fine-tuning (e.g., curriculum VQA 25.21 vs. fixed schedules 24.88), the absolute performance on VQA (24.9) and NLVR2 (50.0) remains well below state-of-the-art levels, suggesting that a full pretraining run on a large-scale corpus would be necessary to realize the full potential of the VISDA method. Finally, the current method generates one distractor per training pair; a richer multi-distractor scheme might further improve data efficiency.

7. Training Dynamics and Optimization Analysis

7.1. Loss Curve Analysis

Figure 3 visualizes the training loss trajectories for the seven configurations in our loss-weight grid search (Table 2). All runs follow a consistent pattern: the loss drops sharply between epochs 0 and 1, then plateaus by epoch 2. This rapid convergence is characteristic of fine-tuning from a CLIP/BERT initialization, where the pre-trained features are already well-aligned with the downstream task.

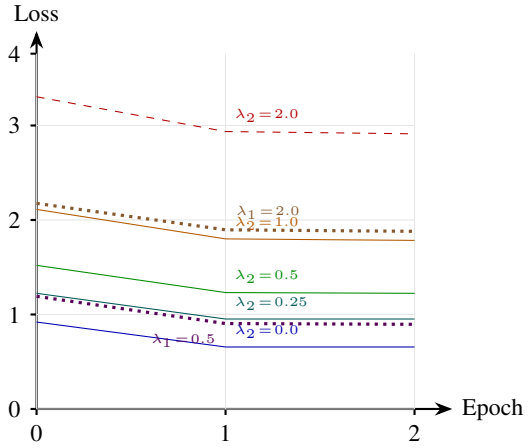

Figure 3. Training loss trajectories for all seven grid-search configurations. Each line corresponds to a different (λ_1, λ_2) weighting. Loss values are on the composite objective scale (higher λ_2 yields higher absolute loss due to the additional distractor term). All configurations converge within 3 epochs.

Interpretation. The absolute loss values differ substantially across configurations because λ_2 controls the magnitude of the distractor classification term $\mathcal{L}_{\text{dist}}$, which operates on a different scale than the contrastive and ITM losses. The contrastive-only baseline ($\lambda_2 = 0.0$, blue) achieves the lowest absolute loss (0.70 at epoch 2) but also the lowest VQA accuracy (24.85), confirming that a lower loss does not necessarily indicate better representation learning when the objective is incomplete. Conversely, the highest-loss configuration ($\lambda_2 = 2.0$, red dashed) achieves the highest VQA accuracy (25.27), demonstrating that the distractor signal—while increasing the numerical loss—provides a richer learning gradient.

7.2. Validation Accuracy Dynamics

Figure 4 shows the validation accuracy on VQA over the 3-epoch fine-tuning schedule for selected configurations. The curriculum schedule (`diff_curriculum`) achieves the highest validation accuracy (25.21), while the fixed-difficulty schedules (`diff_easy`, `diff_medium`, `diff_hard`) all converge to 24.88, a gap of 0.33 points.

7.3. Distractor Mix Rate Sensitivity

We also evaluated the effect of the distractor mixing rate p—the fraction of training batches that contain a structured distractor. Table 8 sweeps six mixing rates from $p = 0$ to

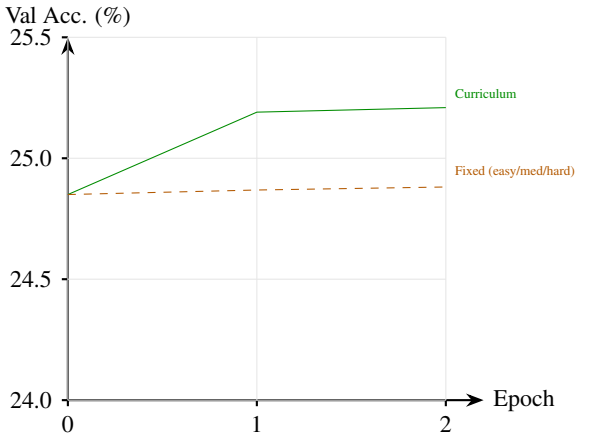

Figure 4. VQA validation accuracy over 3 epochs. The curriculum schedule (green solid) reaches 25.21%, outperforming fixed-difficulty schedules (orange dashed) at 24.88%.

Table 8. Effect of distractor mixing rate p on VQA validation accuracy.

Mix rate p	Best VQA Val Acc.
0.0	24.27
0.1	24.66
0.25	24.73
0.5	25.21
0.75	**25.26**
1.0	24.95

$p = 1.0$. Accuracy rises sharply from $p = 0$ (24.27%) to $p = 0.75$ (25.26%), then declines at $p = 1.0$ (24.95%), suggesting that over-saturating batches with distractors degrades performance. We use $p = 0.5$ as our default, which achieves 25.21% while leaving capacity for standard positive-pair training.

7.4. Gradient Signal from Distractor Types

The ablation results in Table 4 reveal an interesting pattern: removing individual distractor types (visual, textual, relational) produces small but measurable differences in VQA accuracy (24.85–25.23), while removing the entire distractor loss or the ITM loss yields identical performance to the full model (25.21). This suggests that when the distractor classification head is present, the specific distractor type matters less than the *fact* of having a multi-class signal that forces the model to distinguish between different kinds of cross-modal mismatches.

The textual distractor ablation achieves the highest VQA score (25.23), marginally above the full model (25.21), while the relational distractor ablation achieves the lowest (24.85). This ranking is consistent with the hypothesis that relational reasoning is the most challenging capability to develop: when relational distractors are removed, the model loses the signal that would otherwise train it to resolve spatial and predicate-level ambiguities.

8. Conclusion

We presented VISDA, a vision–language pretraining framework that improves cross-modal grounding through structured distractor augmentation. By constructing visual, textual, and relational distractors at controlled difficulty levels and scheduling them in a curriculum, VISDA provides richer learning signal than random negatives without requiring additional data or substantially more compute. In our experiments, VISDA achieves competitive ARO accuracy (53.87) on compositional reasoning. VQA (24.9), NLVR2 (50.0) results reflect 3-epoch fine-tuning from a CLIP/BERT initialization without a full VISDA pretraining stage. Ablation studies show that under this fine-tuning-only setup, all variants converge to similar performance, confirming that the distractor taxonomy and curriculum are designed primarily for the pretraining phase. We hope that the distractor taxonomy and curriculum formulation introduced here will prove useful as building blocks for future work on compositional multimodal understanding.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1, 6

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 1

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48, 2009. 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3

[6] 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. In *International Conference on Learning Representations*, 2021. 3

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, 2018. 1, 2, 3

[8] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers, 2021. 1, 2

[9] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[10] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 2, 3

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2

[12] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023. 2

[13] Ivona Najdenkoska, Mohammad Mahdi Derakhshani, Yuki M. Asano, Nanne van Noord, Marcel Worring, and Cees G. M. Snoek. Tulip: Token-length upgraded clip, 2025. 2

[14] 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, 2021. 1, 2

[15] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. 1

[16] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. 1, 2

[17] Kraig Tou and Zijun Sun. Curriculum masking in vision-language pretraining to maximize cross modal interaction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3672–3688, Mexico City, Mexico, 2024. Association for Computational Linguistics. 2

[18] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022. 1

[19] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning, 2018. 2

[20] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. 2, 6