PANDA: TRAINING-TIME PRIVILEGED ATTRIBUTE SUPERVISION FOR SHIP RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ship Re-Identification requires distinguishing similar vessels across cameras. Identity labels tell the model which images match, but not which visual details matter. Language can provide this guidance, yet sentence-level alignment does not explicitly teach attribute discrimination, while MLLM-based reranking adds inference cost. To address these problems, we propose **P**rivileged **A**ttributes for **N**ative **D**escriptor **A**daptation (**PANDA**), a three-module framework that only uses language supervision during training. Specifically, **P**rivileged **A**ttribute **C**onstruction (**PAC**) extracts shared attributes from paired training views, organizes them into structured descriptions, and creates candidate negatives by replacing one slot with text from another identity. Using these annotations, **F**rozen-LM **A**ttribute **S**upervision (**FAS**) trains visual adapters to generate descriptions and distinguish matching attributes from alternatives through a frozen language model. **A**ttribute-**A**ware **F**usion (**AAF**) then combines identity and semantic features from the attribute-supervised encoder into a single retrieval descriptor. At inference, only the visual encoder and fusion head remain, requiring neither text nor a language model. Experiments on ShipReID-2400 and VesselReID show that PANDA outperforms the compared baselines while maintaining efficient inference, and ablation studies demonstrate the effectiveness of its components.

1 INTRODUCTION

Ship Re-IDentification (Ship-ReID) matches images of the same vessel across cameras (Zhang et al., 2023). Because vessels of the same type often share silhouettes and paint schemes, matching requires finer structural and appearance cues. Existing methods seek these cues through metric learning, part-aware aggregation, and transformer-based modeling (Kalayeh et al., 2018; Wang et al., 2022; He et al., 2021). However, these advances mainly improve feature learning under identity supervision, which specifies which images match but not which visual cues support the match. This gap motivates our question: what additional training supervision can guide an image-only retriever toward these discriminative cues?

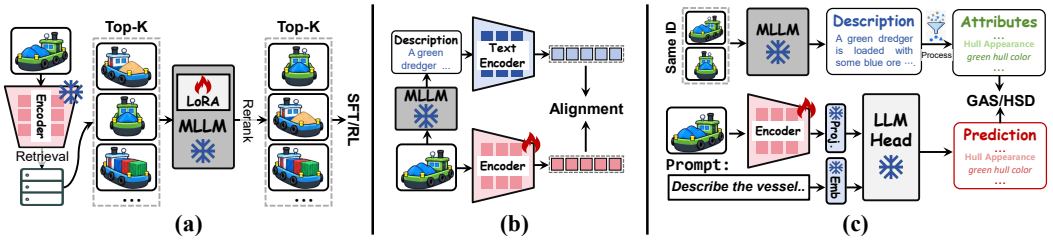


Figure 1: Language-assisted ReID strategies. (a) MLLM reranking uses language-model inference to reorder retrieved candidates. (b) Text–image alignment transfers language information through a shared embedding space. (c) PANDA uses structured attributes and a frozen language model during training, then retrieves images with a visual descriptor.

Language-assisted methods provide several ways to introduce such supervision. Candidate reranking uses an initial retriever to form a shortlist, then applies a Multi-modal Large Language model (MLLM) to reorder candidates (Li & Gong, 2026; Zhang et al., 2026). Since the MLLM only sees this shortlist, missed true matches cannot be recovered, and candidate-level processing brings high

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.

inference cost. Training-time approaches instead exploit language through prompt learning or text–image alignment (Li et al., 2023a; Wang et al., 2026; Gao et al., 2026). These approaches need not retain a language model at deployment, but their embedding-level objectives do not explicitly require predicting individual attribute words or distinguishing competing values of the same attribute.

Motivated by these limitations, we propose a third route: using a frozen language model to supervise attribute generation and discrimination during training. Gradients from these tasks update the visual encoder, while the language model is removed at inference. Developing this route raises three questions: how to construct useful attribute targets, how to teach the encoder to distinguish similar attribute values, and how to combine the learned visual features into the encoded descriptor.

We address these questions with **P**rivileged **A**ttributes for **N**ative **D**escriptor **A**daptation (**PANDA**), which comprises three modules. First, **P**rivileged **A**ttribute **C**onstruction (**PAC**) constructs attribute targets by using a frozen annotator to extract shared cues from same-identity, cross-camera training pairs. Deterministic rules organize these cues into six slots, replaceing one slot with text from a similar training identity yields a candidate negative. Second, **F**rozen-LM **A**ttribute **S**upervision (**FAS**) trains visual adapters to predict attributes and distinguish similar attribute values through Generative Attribute Supervision (GAS) and Hard-Negative Slot Discrimination (HSD). GAS predicts structured descriptions, whereas HSD selects between original and donor-slot candidates. The language model remains frozen throughout both stages. Third, **A**ttribute-**A**ware **F**usion (**AAF**) converts the supervised visual features into a retrieval descriptor by combining global features with learnable identity and semantic query readouts. Entropy-modulated interaction uses the semantic readouts to refine the identity representation, and both branches form the final descriptor. The semantic queries are learned jointly with the ReID objectives. Inference retains only the visual encoder and retrieval head. Extensive experiments on ShipReID-2400 and VesselReID show that PANDA outperforms the compared baselines, with ablations supporting the effectiveness of its components. Our contributions are summarized as follows:

- We develop PANDA, a training-time language-supervision framework for Ship ReID that updates visual adapters through a frozen language model and requires no language inference at deployment.
- We introduce paired-view privileged attribute construction, frozen-language-model supervision that transfers the constructed attributes through token-level generation and slot-based counterfactual discrimination, and an attribute-aware fusion head that produces descriptor.
- Extensive experiments on two ship ReID benchmarks show that PANDA outperforms existing visual and language-enhanced ReID methods, while maintaining efficient inference. Further analyses validate the contribution of its supervision stages and descriptor components.

2 Related Work

2.1 Ship Re-Identification

Ship Re-Identification distinguishes visually similar vessels using fine-grained, cross-view-stable cues. Existing ReID and ship ReID methods mainly improve visual descriptors by better metric objectives, stronger backbones, or local/part-aware aggregation (He et al., 2021; Sun et al., 2021; Huang et al., 2020; Liu et al., 2025; Wang et al., 2025). *However, these methods mostly change how visual evidence is encoded, while identity labels remain the primary supervision. PANDA instead focuses on what supervision should shape the descriptor: it explicitly uses stable vessel attributes to guide the visual encoder toward identity-preserving cues.*

2.2 Language-enhanced Re-Identification

Language–enhanced ReID methods commonly rely on text–image alignment or MLLM-based instruction following and reranking (Li et al., 2023a; He et al., 2024; Lu et al., 2025; Wang et al., 2026; Zhang et al., 2026; Li & Gong, 2026). *However, alignment-based methods often provide only holistic language supervision, while MLLM-based methods introduce inference–time language reasoning and candidate-recall dependence. PANDA instead uses a frozen language model only during training for token-level attribute supervision and keeps inference as purely visual retrieval.*

2.3 Learning with Privileged Information

Learning with privileged information improves test-time models using signals available only during training (Vapnik & Vashist, 2009). This principle has been widely used in distillation and multi-modal learning to enrich supervision without increasing inference cost (Gou et al., 2021; Lopez-Paz et al., 2015; Hoffman et al., 2016). *PANDA adopts this idea by treating vessel attributes as privileged supervision, while removing both attributes and the language model at inference.*

3 Methodology

3.1 Problem Formulation

Given a labeled training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M}$, ship Re-identification aims to learn a model that retrieves gallery images sharing the same vessel identity as a query image. At test time, the model ranks a gallery set $\mathcal{G}$ for each query x_q.

3.2 Framework Overview

PANDA treats vessel attributes as training-only privileged information for ship ReID. The framework follows an attribute generation $\rightarrow$ attribute learning pipeline. First, Privileged Attribute Construction (PAC) generates reliable supervision offline: a frozen annotator MLLM compares cross-camera images of the same vessel, keeps only view-stable attributes, organizes them into semantic slots, and constructs single-slot counterfactuals from similar identities (Figure 2). Second, the generated attributes are transferred to the visual model. Frozen-LM Attribute Supervision (FAS) inserts visual tokens into a frozen language model and trains them with Generative Attribute Supervision (GAS) and Hard-Negative Slot Discrimination (HSD), encouraging the tokens to both generate true attributes and reject counterfactual ones. Attribute-Aware Fusion (AAF) then removes the language model and builds a visual descriptor that retains identity and attribute evidence (Fig. 3). At inference, PANDA performs nearest-neighbor retrieval with this descriptor.

3.3 Privileged Attribute Construction

As shown in Figure 2, Privileged Attribute Construction (PAC) converts training images into structured privileged supervision through two steps: cross-view attribute generation and single-slot counterfactual construction.

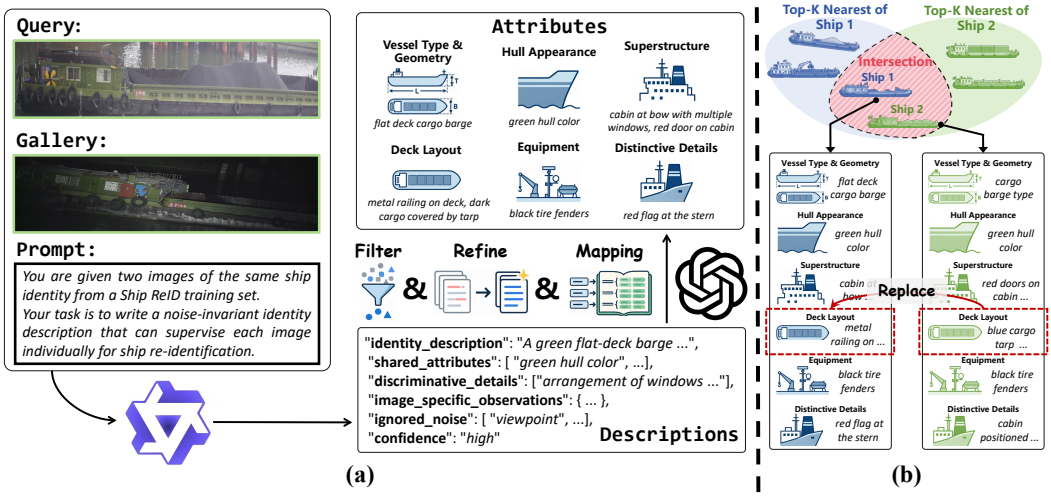


Figure 2: Overview of Privileged Attribute Construction. (a) Cross-view Attribute Generation extracts cross-view-stable vessel attributes. (b) Single-slot Counterfactual Construction builds hard negatives by replacing one attribute slot.

Cross-view attribute generation. Attributes from a single image may reflect viewpoint, background, or illumination. To obtain identity-preserving cues, we sample cross-camera image pairs

for each vessel identity and ask a frozen annotator MLLM to describe only attributes consistently visible in both views. The shared attributes are then organized into K semantic slots, such as vessel geometry, hull appearance, superstructure, deck layout, equipment, and distinctive details. For image x_i, PAC produces

$$Y_i = \text{Serialize}\big(S_i^{(1)}, \ldots, S_i^{(K)}\big), \tag{1}$$

where $S_i^{(k)}$ denotes the attributes assigned to the k-th slot. This slot structure provides a stable format for language supervision and separates different types of visual evidence.

Single-slot counterfactual construction. To obtain discriminative supervision, PAC further constructs counterfactual attributes. For each record, we select a similar identity as the donor and replace one non-empty slot while keeping the remaining slots unchanged. This yields a true slot c_i^+ from the original vessel and a counterfactual slot c_i^- from the donor identity. Since the negative differs by only one real attribute slot, it encourages the model to distinguish fine-grained identity cues beyond coarse vessel appearance. Further details are provided in Appendix C.1.

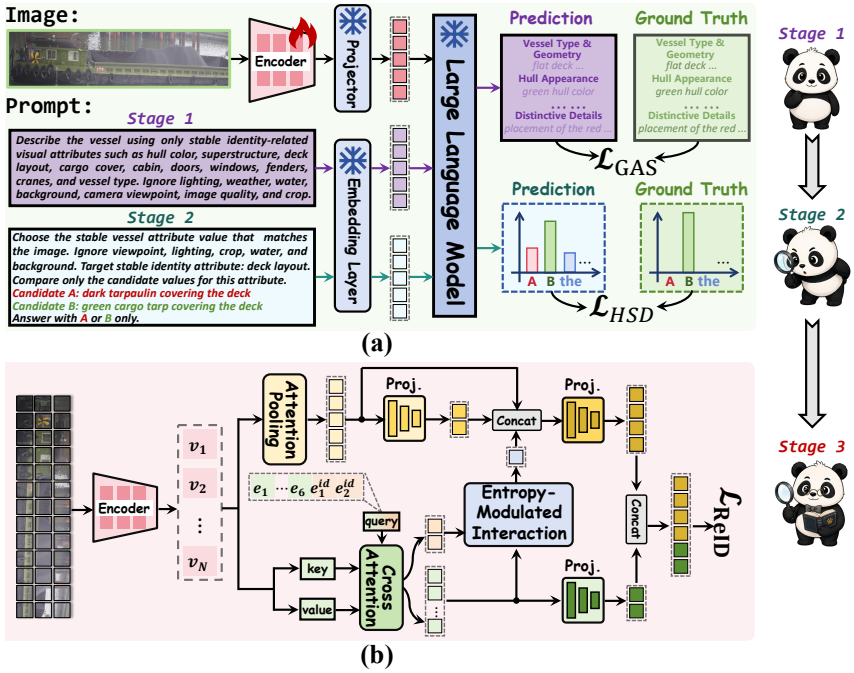


Figure 3: Training process of PANDA. (a) Frozen language-model supervision uses the slot-structured attributes produced by privileged attribute construction to supervise visual tokens. The visual encoder is trained to generate the attribute description and to reject single-slot counterfactuals while the language model remains frozen. (b) Attribute-Aware fusion converts the supervised visual tokens into an image-only retrieval descriptor. It combines global visual evidence, identity tokens, and semantic query tokens through entropy-modulated interaction.

3.4 Frozen-LM Attribute Supervision

As shown in Figure 3(a), Frozen-LM Attribute Supervision transfers the generated attributes to the visual encoder through a frozen language model in two consecutive stages: Generative Attribute Supervision and Hard-Negative Slot Discrimination. Given an image x_i, the visual encoder produces visual tokens

$$T_i = E_\theta(x_i) \in \mathbb{R}^{N \times d}. \tag{2}$$

The tokens are inserted into the frozen language model with parameters ϕ, and gradients are back-propagated only to the visual encoder while ϕ remains fixed.

Generative attribute supervision. The first stage trains the visual tokens to support generation of the structured attribute description. Given a prompt P and tokenized attributes $Y_i =$

$(y_{i,1}, \ldots, y_{i,L_i})$, we use the token-level loss as follows:

$$\mathcal{L}_{\mathrm{GAS}} = -\frac{1}{\sum_i L_i} \sum_i \sum_{t=1}^{L_i} \log p_\phi\big(y_{i,t} \mid P, T_i, y_{i,<t}\big). \tag{3}$$

Unlike sentence-level image-text alignment, this objective supervises individual semantic tokens and directly shapes the visual representation.

Hard-negative slot discrimination. The second stage builds on this generative supervision by training the visual tokens to distinguish true slots from single-slot counterfactuals. Both candidates in each PAC pair (c_i^+, c_i^-) are included in a discriminative prompt conditioned on T_i, with the two A/B orderings sampled equiprobably. Let τ_i denote the correct option and $z_{i,o}$ the frozen LM's next-token logit for answer-label token $o \in \{\mathrm{A}, \mathrm{B}\}$, extracted at the final non-padding position of the complete multimodal prompt. The candidates are input context rather than generation targets, no answer token is supplied. The per-example classification loss and weighted batch objective are

$$\begin{aligned} \ell_i^{\mathrm{CF}} &= -\log \frac{\exp z_{i,\tau_i}}{\exp z_{i,\mathrm{A}} + \exp z_{i,\mathrm{B}}}, \\ \mathcal{L}_{\mathrm{HSD}} &= \frac{\sum_{i \in \mathcal{B}_{\mathrm{CF}}} w_i \ell_i^{\mathrm{CF}}}{\sum_{i \in \mathcal{B}_{\mathrm{CF}}} w_i}. \end{aligned} \tag{4}$$

where $\mathcal{B}_{\mathrm{CF}}$ contains valid examples, and w_i combines PAC sample weighting with target-slot frequency correction. Equation 4 applies when the total weight is positive. Weight definitions and boundary cases appear in Appendix C.2. To retain the generative supervision of equation 3, we include a low-weight replay term:

$$\mathcal{L}_{\mathrm{stage2}} = \mathcal{L}_{\mathrm{HSD}} + \lambda_{\mathrm{r}} \cdot \mathcal{L}_{\mathrm{GAS}}. \tag{5}$$

3.5 Attribute-Aware Fusion

After FAS, the language model is removed. As shown in Fig. 3(b), Attribute-Aware Fusion (AAF) converts the supervised visualtokens into a descriptor. It uses a global stream for holistic visual evidence, an identity stream for instance-level cues, and a semantic stream for complementary attribute-informed evidence. Given visual tokens $T_i = \{v_1, \ldots, v_N\}$, AAF extracts a global feature, K semantic query tokens , and R identity tokens:

$$\begin{aligned} g_i &= \mathrm{AttentionPooling}(T_i), \\ (S_i, U_i) &= \mathrm{CrossAttention}\big([Q_{\mathrm{sem}}, Q_{\mathrm{id}}], T_i\big), \\ S_i &= \{s_i^k\}_{k=1}^K, \qquad U_i = \{u_i^r\}_{r=1}^R. \end{aligned} \tag{6}$$

We use $K = 6$ semantic queries and $R = 2$ identity queries.

$$C_i = \mathrm{Attention}(U_i, S_i), \qquad \Delta_i = \mathrm{EMI}(U_i, C_i, \mathcal{H}), \tag{7}$$

where C_i is the attended semantic context, $\mathcal{H}$ denotes the attention entropy, EMI uses normalized attention entropy to adaptively scale the semantic correction, giving larger weight to confident, concentrated interactions while suppressing diffuse ones, its detailed formulation is provided in Appendix C.3.The identity branch combines the global feature, identity tokens, and the semantic-guided residual. The semantic branch retains multiple semantic readouts in separate coordinates:

$$\begin{aligned} a_i &= \mathrm{Proj}_{\mathrm{id}}\big([\mathrm{Proj}_g(g_i + \Delta_i), \mathrm{Proj}_u([u_i^1, \ldots, u_i^R]),]\big), \\ b_i &= \frac{1}{\sqrt{K}} \left[\mathrm{norm}(\mathrm{Proj}_{\mathrm{sem}}(s_i^1)), \ldots, \mathrm{norm}(\mathrm{Proj}_{\mathrm{sem}}(s_i^K))\right]. \end{aligned} \tag{8}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. The final descriptor concatenates the normalized identity and attribute branches:

$$f_i = \left[\sqrt{\lambda_{\mathrm{id}}}\,\mathrm{norm}(a_i), \sqrt{\lambda_{\mathrm{sem}}}\,\mathrm{norm}(b_i)\right], \qquad \lambda_{\mathrm{id}} + \lambda_{\mathrm{sem}} = 1, \tag{9}$$

The separate query blocks retain multiple learned readouts without imposing fixed attribute meanings.

Optimization and inference. AAF is trained with identity classification, triplet, and ranking losses:

$$\mathcal{L}_{\text{ReID}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}} + \lambda_{\text{r}} \cdot \mathcal{L}_{\text{rank}}. \tag{10}$$

The three terms are defined as

$$\begin{aligned}
\mathcal{L}_{\text{ce}} &= -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(W_{y_i}^{\top} f_i)}{\sum_{c=1}^{C} \exp(W_c^{\top} f_i)}, \\
\mathcal{L}_{\text{tri}} &= \frac{1}{B}\sum_{i=1}^{B} \left[m + d(f_i, f_i^{+}) - d(f_i, f_i^{-})\right]_{+}, \\
\mathcal{L}_{\text{rank}} &= -\frac{1}{B}\sum_{i=1}^{B} \frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \log \frac{\exp(f_i^{\top} f_j/\tau)}{\sum_{l \neq i} \exp(f_i^{\top} f_l/\tau)}.
\end{aligned} \tag{11}$$

where B is the batch size, W_c is the classifier weight for identity c, $d(\cdot,\cdot)$ denotes descriptor distance, f_i^{+} and f_i^{-} are the hardest positive and negative samples in the batch, and $\mathcal{P}_i = \{j \neq i \mid y_j = y_i\}$. During inference, PAC and FAS are discarded, PANDA computes $f(x)$ with AAF and ranks gallery images by nearest-neighbor search.

4 Experiments

4.1 Experiment Setup

Datasets We evaluate PANDA on two Ship Re-Identification benchmarks. ShipReID-2400 (Liu et al., 2025) contains 17,241 images of 2,400 ship identities collected from a real-world waterway monitoring system over 53 months. Following the official split, 1,800 identities are used for training, 300 for validation, and 300 for testing, the query sets contain 691 validation queries and 706 test queries, with 4,253 gallery images. VesselReID (Zhang et al., 2023) contains 30,587 images of 1,248 vessel identities captured under different viewpoints, places, times, and weather conditions. We follow the standard split protocol of each dataset and report performance on the test set.

Comparison Baselines We compare PANDA with three categories of methods. First, we include generic visual ReID baselines that learn image-only descriptors through metric learning, transformer backbones, part-aware modeling, or object-level interaction, including DPM (Hermans et al., 2017), TransReID (He et al., 2021), DC-Former (Li et al., 2023b), AdaSP (Zhou et al., 2023), MSINet (Gu et al., 2023), and PFDNet (Wang et al., 2022). Second, we compare with vehicle- and ship-oriented methods, including FED (Qiao et al., 2020), UFDN (Qian et al., 2022), HRCN (Zhao et al., 2021), and the ship-specific D2InterNet (Liu et al., 2025). Third, we compare with language-enhanced or MLLM-based ReID methods, including CLIP-ReID (Li et al., 2023a), Instruct-ReID (He et al., 2024), LVLM-ReID (Wang et al., 2026), MUSE (Li & Gong, 2026), and ReID-R (Zhang et al., 2026). These baselines cover standard visual descriptor learning, domain-oriented vessel/vehicle retrieval, and recent language-assisted ReID. More details on the baseline categories and implementation choices are provided in Appendix D.

Implementation Details PANDA is trained in three stages: Generative Attribute Supervision, Hard-negative Slot Discrimination, and ReID training with Attribute-Aware Fusion. The first two stages update only visual LoRA adapters through a frozen language model. While the final stage removes the language model and trains the adapters and retrieval head with ReID losses. The base visual weights remain frozen throughout. The visual encoder and frozen annotator are based on the Qwen3-VL model (Bai et al., 2025a), with privileged attributes generated exclusively from training images. Inference retains only the visual encoder and retrieval head. We report mAP and Rank-1/5/10 for retrieval accuracy, and GFLOPs and single-image inference latency for efficiency. Detailed model configurations, metrics, and training settings are provided in Appendix D.

4.2 Performance Comparison

Table 1 compares PANDA with general visual ReID methods, methods designed for vehicle or ship ReID, and language-assisted approaches. PANDA achieves the highest scores across all reported

metrics on both datasets. On ShipReID-2400, it surpasses the ship-specific baseline D2InterNet by 9.59 % in mAP and 13.63 % in Rank-1. It also outperforms the language-based reranking method ReID-R while requiring no language-model inference.

Table 1: Performance Comparison on ShipReID-2400 and VesselReID datasets.

Datasets	**ShipReID-2400**				**VesselReID**			
Method	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
DPM	20.20	23.50	37.80	44.90	56.25	68.55	86.00	91.65
FED	28.16	33.14	47.17	56.94	55.93	68.49	87.07	91.96
AdaSP	32.14	36.83	56.80	64.73	61.35	74.26	90.14	94.04
MSINet	32.60	37.00	54.80	64.30	64.39	74.70	90.71	94.73
DC-Former	25.20	26.90	43.80	52.50	59.39	70.94	88.51	92.40
TransReID	29.80	34.10	48.70	57.50	60.56	70.62	89.27	93.41
UFDN	30.63	35.13	49.29	58.92	55.19	67.80	85.88	91.78
HRCN	25.68	27.90	48.30	56.94	55.91	67.17	86.82	92.59
PFDNet	23.60	26.90	42.10	49.00	51.71	64.16	82.61	88.26
D2InterNet	41.60	46.00	62.60	70.00	61.70	72.82	89.01	92.91
Instruct-ReID	25.30	27.90	46.20	55.00	52.40	64.60	84.93	89.83
CLIP-ReID	32.50	36.40	54.10	61.20	60.49	71.50	87.76	93.22
LVLM-ReID	24.90	26.91	45.75	55.24	52.19	63.59	83.18	89.01
MUSE	41.96	46.96	62.93	70.02	63.13	76.69	89.83	93.01
ReID-R	42.80	48.16	63.93	70.96	64.11	80.44	91.15	92.91
PANDA (Ours)	**51.19**	**59.63**	**72.38**	**77.90**	**76.02**	**88.20**	**95.86**	**97.18**

On VesselReID, PANDA exceeds ReID-R by 11.91 % in mAP and 7.76 % in Rank-1. Several baselines already achieve high Rank-5 and Rank-10 accuracy, meaning that correct matches often appear among their leading candidates. PANDA's gains in mAP and Rank-1 indicate improved ordering of these matches, particularly at the first retrieval position.

Together, these results support the effectiveness of PANDA's training-time language supervision and visual descriptor learning. Rather than applying language reasoning to an initial shortlist, PANDA uses attribute supervision to train the visual encoder and ranks the full gallery with the resulting descriptors. This yields strong retrieval performance without retaining a language model or depending on a separate retriever's candidate pool.

4.3 Ablation Study

We ablate Frozen-LM Attribute Supervision and Attribute-Aware Fusion in the main text to assess the contributions of attribute supervision and descriptor construction. Additional experiments examine Privileged Attribute Construction, alternative MLLM backbones, annotator capacity, and interaction with inference-time MLLM reranking in Appendices A.1, A.2, A.3 and A.4.

Ablation on Frozen-LM Attribute Supervision Table 2 evaluates FAS through controlled supervision baselines and component ablations, complementing the cross-method comparison in Table 1 by holding the visual backbone and downstream ReID setup fixed. Removing FAS reduces performance close to Pure Visual ReID, highlighting the importance of the language-based supervision stage. Direct Attribute Supervision predicts PAC attributes with multi-label classification, while Holistic Description SFT uses the original full descriptions as generative targets. Both improve substantially over Pure Visual ReID but remain clearly below full PANDA, showing that neither direct attribute prediction nor holistic language generation alone fully captures the benefit of FAS.

The component ablations further show complementary roles of GAS and HSD. Removing GAS consistently degrades retrieval performance, supporting the importance of learning stable attribute semantics. Removing HSD has a modest effect on ShipReID-2400 but causes a clearer degradation on VesselReID, indicating that counterfactual discrimination provides additional fine-grained supervision. Combining both objectives achieves the strongest overall mAP and Rank-1 performance on both datasets.

Table 2: Ablation and controlled comparison of Frozen-LM Attribute Supervision.

Datasets	ShipReID-2400				VesselReID			
Setting	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
Full	**51.19**	**59.63**	72.38	**77.90**	**76.02**	**88.20**	**95.86**	**97.18**
FAS Component Ablations								
w/o GAS	48.24	54.67	70.25	76.77	73.09	83.94	88.85	90.86
w/o HSD	51.04	58.92	**73.09**	77.48	72.02	83.70	89.29	90.05
w/o FAS	39.23	47.67	63.68	66.06	67.00	76.33	91.15	93.23
Controlled Supervision Baselines								
Pure Visual ReID	39.24	46.65	60.37	65.90	66.68	75.74	90.97	93.74
Direct Attribute Supervision	46.50	52.90	69.70	76.63	72.33	81.88	86.52	90.27
Holistic Description SFT	46.52	53.53	70.76	75.59	72.88	82.14	90.92	91.92

Ablation on Attribute-Aware Fusion Table 3 evaluates three AAF readouts: global features (G), the interaction-enhanced identity anchor (A), and semantic query features (S). These settings vary the outputs included in the descriptor, the A readout still uses semantic tokens when S is excluded. Single-readout performance varies across datasets. On ShipReID-2400, G and A achieve similar mAP (45.86% and 45.69%), while S reaches 42.01%. On VesselReID, however, S performs best among the three, reaching 70.61% mAP.

Table 3: Ablation of AAF descriptor readouts. G, A, and S denote global features, the interaction-enhanced identity anchor, and semantic query features, respectively.

Components			ShipReID-2400				VesselReID			
G	A	S	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
✓			45.86	54.53	67.85	73.51	66.16	76.69	90.29	95.24
	✓		45.69	53.12	68.27	73.65	63.50	75.25	90.79	96.30
		✓	42.01	47.03	62.61	68.41	70.61	83.62	92.47	95.61
✓	✓		50.27	57.79	72.66	76.91	73.42	85.37	93.91	96.48
✓		✓	50.50	56.52	72.80	78.47	73.73	86.20	95.67	97.30
	✓	✓	45.39	52.12	68.27	74.36	70.40	85.44	93.72	95.92
✓	✓	✓	**51.19**	**59.63**	**72.38**	**77.90**	**76.02**	**88.20**	**95.86**	**97.18**

Combining the readouts improves retrieval. Adding S to G increases mAP from 45.86% to 50.50% on ShipReID-2400 and from 66.16% to 73.73% on VesselReID. Combining G with A raises ShipReID-2400 Rank-1 from 54.53% to 57.79%. The full $G + A + S$ descriptor achieves the highest mAP and Rank-1 on both datasets: 51.19%/59.63% on ShipReID-2400 and 76.02%/88.20% on VesselReID. However, $G + S$ retains higher Rank-5/10 on ShipReID-2400 and Rank-10 on VesselReID. These results support combining global, identity-anchor, and semantic readouts to improve overall ranking and top-1 matching, without implying gains at every retrieval rank.

We further examine the interaction mechanism within AAF. As shown in Table 4, removing EMI degrades performance, confirming the benefit of semantic–identity interaction. Removing the entropy-dependent scaling causes a smaller but consistent drop on both datasets, showing that the attention entropy $\mathcal{H}$ provides additional benefit by adaptively modulating the semantic correction.

Table 4: Ablation of the interaction mechanism in Attribute-Aware Fusion.

Datasets	ShipReID-2400				VesselReID			
Setting	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
w/o EMI	50.06	56.94	72.10	77.48	73.23	86.09	94.46	96.98
w/o Entropy Scaling ($\mathcal{H}$)	50.36	58.07	72.09	76.32	75.26	86.96	95.07	97.13
Full AAF	**51.19**	**59.63**	**72.38**	**77.90**	**76.02**	**88.20**	**95.86**	**97.18**

4.4 Further Analysis

We further analyze PANDA from the perspectives of robustness and efficiency. In the main paper, we study the sensitivity to the replay weight λ_r and compare the accuracy–efficiency trade-off across methods. Qualitative analysis provided in Appendix B, including feature-space visualization, retrieval case studies, and attention visualization, which further illustrate how PANDA improves identity separation, avoids candidate-recall limitations, and attends to fine-grained vessel cues.

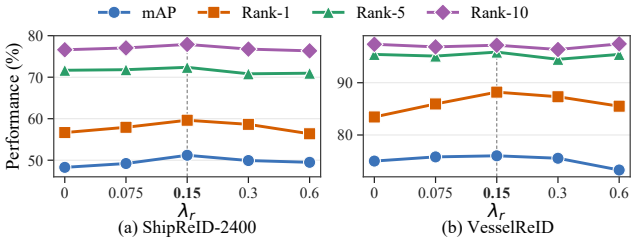


Figure 4: Sensitivity analysis of λ_r.

Hyperparameter Sensitivity Analysis We study the sensitivity of PANDA to the replay weight λ_r in the second frozen-language-model supervision stage. This parameter controls the strength of generative attribute replay when optimizing hard-negative slot discrimination. As shown in Figure 4, PANDA is stable across a wide range of λ_r values on both ShipReID-2400 and VesselReID, indicating that the method is not sensitive to precise tuning of this hyperparameter.

A moderate replay weight gives the best overall performance. When $\lambda_r = 0$, the model relies only on hard-negative slot discrimination in the second stage, which weakens the generative attribute knowledge learned in the first stage. Increasing λ_r to 0.15 improves performance on both datasets, suggesting that retaining a small amount of generative supervision helps preserve stable attribute semantics while learning fine-grained counterfactual discrimination. Larger values do not bring further gains and may slightly reduce performance, likely because excessive replay weakens the discriminative focus of the hard-negative objective.

Efficiency Analysis Figure 5 compares retrieval performance with computational cost, measured by FLOPs and latency. PANDA lies on the Pareto frontier on both ShipReID-2400 and VesselReID, achieving the best retrieval accuracy while maintaining competitive inference cost. This supports the main design claim of PANDA: attribute knowledge can be transferred into the descriptor during training rather than invoked through expensive language-model reasoning at inference. As a result, PANDA achieves a better accuracy–efficiency trade-off than both purely visual baselines and inference-time language-enhanced methods.

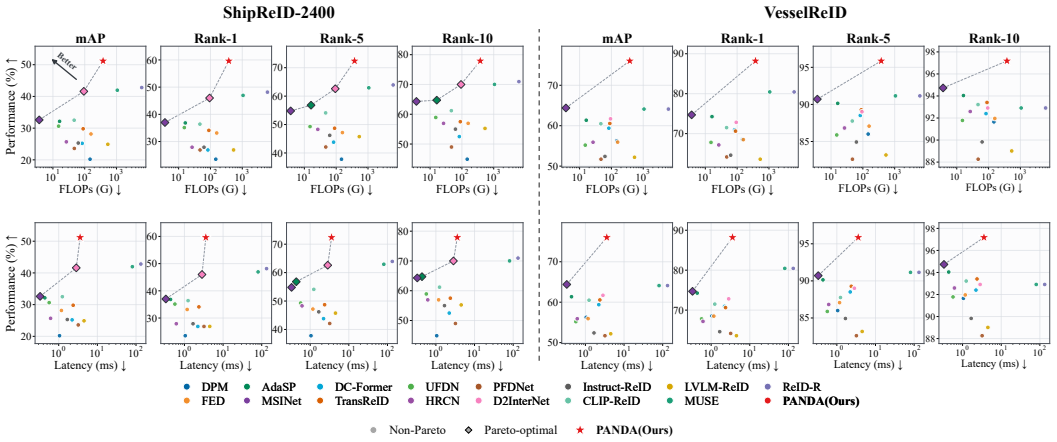


Figure 5: Efficiency–performance comparison.

5 Conclusion

This paper explored a training-time use of language models for ship re-identification: rather than keeping language models in the retrieval loop or aligning images to holistic text embeddings, we use a frozen language model as a supervision interface for learning visual descriptors. The proposed PANDA framework constructs cross-view-stable vessel attributes, transfers them to visual representations through frozen language-model supervision, and integrates the resulting identity and semantic evidence with Attribute-Aware Fusion. This design keeps deployment fully image-only while benefiting from fine-grained attribute knowledge during training. Experiments on ShipReID-2400 and VesselReID show that PANDA improves over visual, ship-oriented, and language-enhanced baselines, and achieves a favorable accuracy–efficiency trade-off. We hope this perspective encourages broader use of MLLMs as training-time semantic supervisors for efficient visual retrieval.

AI USE STATEMENT

We used generative AI tools in two ways. First, generative AI is part of the proposed method: a frozen multimodal language model is used to construct training-only privileged vessel attributes from cross-view image pairs, and a frozen language model is used during training to provide attribute supervision. Second, we used generative AI tools for manuscript editing, code/documentation organization, and experiment-report preparation.

We did not use generative AI tools to fabricate experimental results, alter evaluation metrics, or replace author responsibility for scientific claims. All AI-assisted text, code, annotations, and analyses were reviewed by the authors. Experimental results were produced by running the described models and evaluation protocols. The authors take full responsibility for the final content of this work, including all claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies ship re-identification on public vessel/ship ReID benchmarks. The experiments do not involve human subjects or personally identifiable human information. However, ship re-identification is related to visual retrieval and maritime monitoring, and such technologies may raise privacy, tracking, or dual-use concerns depending on deployment context. Our study is intended for research on robust visual retrieval and does not release new surveillance data. We encourage responsible use of vessel re-identification systems with appropriate legal, safety, and privacy safeguards.

REPRODUCIBILITY STATEMENT

We provide the main algorithmic details of PANDA in Sec. 3, including privileged attribute construction, frozen-language-model supervision, and attribute-aware fusion. Additional implementation details, prompt templates, slot definitions, training pipeline, evaluation metrics, and experimental settings are provided in Appendices C and D. We use public datasets and follow their standard evaluation protocols. Code and processed annotation files will be released to support reproducibility.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Ya Gao, Shihao Li, Zhaojun Liu, Aihua Zheng, Chenglong Li, and Jin Tang. Chain-of-thought guided multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37705–37714, 2026.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19243–19253, 2023. doi: 10.1109/CVPR52729.2023.01844.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14993–15002, 2021. doi: 10.1109/ICCV48922.2021.01474.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, Donglian Qi, and Yunfeng Yan. Instruct-reid: A multi-purpose person re-identification task with instructions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17521–17531, 2024. doi: 10.1109/CVPR52733.2024.01659.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Houjing Huang, Wenjie Yang, Jinbin Lin, Guan Huang, Jiamiao Xu, Guoli Wang, Xiaotang Chen, and Kaiqi Huang. Improve person re-identification with part awareness learning. *IEEE Transactions on Image Processing*, 29:7468–7481, 2020. doi: 10.1109/TIP.2020.3003442.

Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1062–1071, 2018. doi: 10.1109/CVPR.2018.00117.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025. URL https://mlanthology.org/tmlr/2025/li2025tmlr-llavaonevision/.

Jiachen Li and Xiaojin Gong. Multimodal llm-empowered re-ranking for generalizable person re-identification. *arXiv preprint arXiv:2606.16161*, 2026.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1405–1413, 2023a.

Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1415–1423, 2023b.

Baolong Liu, Roukai Huang, Xin Pan, Chuanhuang Li, Jie Sun, Jianfeng Dong, and Xun Wang. Advancing ship re-identification in the wild: The shipreid-2400 benchmark dataset and d2internet baseline method. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 106–115, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3729892. URL https://doi.org/10.1145/3726302.3729892.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: selective multi-image questioner for interactive person re-identification. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-25.

Wen Qian, Hao Luo, Silong Peng, Fan Wang, Chen Chen, and Hao Li. Unstructured feature decoupling for vehicle re-identification. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 336–353, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19781-9.

Dalei Qiao, Guangzhong Liu, Feng Dong, She-Xiang Jiang, and Likun Dai. Marine vessel re-identification: A large-scale dataset and global-and-local fusion-based discriminative feature learning. *IEEE Access*, 8:27744–27756, 2020.

Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):902–917, 2021. doi: 10.1109/TPAMI.2019.2938523.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130. Advances in Neural Networks Research: IJCNN2009.

Han Wang, Shengyang Li, Jian Yang, Yuxuan Liu, Yixuan Lv, and Zhuang Zhou. Cross-modal ship re-identification via optical and sar imagery: A novel dataset and method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7873–7883, October 2025.

Qizao Wang, Bin Li, and Xiangyang Xue. When large vision-language models meet person re-identification. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10497–10501, 2026. doi: 10.1109/ICASSP55912.2026.11460593.

Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2540–2549, 2022.

Qian Zhang, Mingxin Zhang, Jinghe Liu, Xuanyu He, Ran Song, and Wei Zhang. Unsupervised maritime vessel re-identification with multi-level contrastive learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):5406–5418, 2023. doi: 10.1109/TITS.2023.3243591.

Quan Zhang, Jingze Wu, Jialong Wang, Xiaohua Xie, Jianhuang Lai, and Hongbo Chen. Thinking before matching: A reinforcement reasoning paradigm towards general person re-identification. *arXiv preprint arXiv:2604.19218*, 2026.

Jiajian Zhao, Yifan Zhao, Jia Li, Ke Yan, and Yonghong Tian. Heterogeneous relational complement for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 205–214, October 2021.

Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19691–19701, June 2023.

APPENDIX CONTENTS

A **Additional Experiments** 14
A.1 Effectiveness of Privileged Attribute Construction 14
A.2 Generalization Across Multi-modal Large Language Models 14
A.3 Effect of Annotator Capacity 14
A.4 Interaction with Inference-Time MLLM Reranking 15
A.5 Visual Grounding of Hard-Negative Slot Discrimination 15
A.6 Quality Audit of PAC Annotations 15

B **Qualitative Analysis** 17
B.1 Feature Space Visualization 17
B.2 Retrieval Case Study 17
B.3 Attention Visualization 19

C **Additional Method Details** 19
C.1 Privileged Attribute Construction 19
C.2 Frozen-LM Attribute Supervision 21
C.3 Attribute-Aware Fusion 22
C.4 Training Pipeline 24

D **Detailed Experimental Setup** 24
D.1 Datasets 24
D.2 Evaluation Metrics 25
D.3 Comparison Baselines 25
D.4 Implementation Details 25
D.5 Direct Attribute Supervision 26
D.6 Holistic Description SFT 26

A Additional Experiments

A.1 Effectiveness of Privileged Attribute Construction

Table 5 evaluates whether the attributes produced by Privileged Attribute Construction provide better supervision for language-based ReID methods. We replace the original annotations used by CLIP-ReID (Li et al., 2023a), Instruct-ReID (He et al., 2024), and LVLM-ReID (Wang et al., 2026) with the slot-structured PAC attributes, while keeping the corresponding training framework unchanged.

Table 5: Performance comparison using different annotation strategies.

Datasets		ShipReID-2400				VesselReID			
Method	**Annotation**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
CLIP-ReID	Origin	32.50	36.40	54.10	61.20	60.49	71.50	87.76	**93.22**
	PAC	**33.88**	**39.85**	**56.71**	**63.06**	**60.74**	**71.63**	**88.26**	93.16
Instruct-ReID	Origin	25.30	27.90	46.20	55.00	52.40	64.60	84.93	**89.83**
	PAC	**26.49**	**28.51**	**49.08**	**56.70**	**53.80**	**66.90**	**86.24**	89.27
LVLM-ReID	Origin	24.90	26.91	45.75	55.24	52.19	63.59	83.18	89.01
	PAC	**27.43**	**30.74**	**47.03**	**56.09**	**53.30**	**65.41**	**84.43**	**90.27**

Across all three methods, PAC attributes consistently improve performance on ShipReID-2400. For CLIP-ReID, mAP increases from 32.50% to 33.88% and Rank-1 from 36.40% to 39.85%. Similar gains are observed for Instruct-ReID and LVLM-ReID, with LVLM-ReID improving from 24.90% to 27.43% mAP and from 26.91% to 30.74% Rank-1. These results show that cross-view-stable, slot-structured attributes provide cleaner and more discriminative language supervision than the original annotations, especially on the noisier ShipReID-2400 benchmark.

On VesselReID, PAC attributes also improve most metrics, although the gains are generally smaller. This is consistent with the observation that VesselReID has less severe visual noise and stronger coarse retrieval performance. Overall, the results validate Privileged Attribute Construction as a useful annotation strategy beyond PANDA itself: even when plugged into existing language-alignment methods, the structured attributes improve retrieval performance.

A.2 Generalization Across Multi-modal Large Language Models

Table 6 evaluates PANDA with different MLLMs. PANDA achieves strong performance across all three models, indicating that its training-time attribute supervision and attribute-aware descriptor are not tied to a specific model architecture. On ShipReID-2400, InternVL-8B performs best, reaching 54.46% mAP and 60.62% Rank-1, while LLaVA-OneVision-7B (Li et al., 2025) and Qwen3-VL-2B (Bai et al., 2025a) also remain competitive. On VesselReID, InternVL-8B (Chen et al., 2024) obtains the best mAP, Rank-5, and Rank-10, whereas Qwen3-VL-2B achieves the highest Rank-1. These results show that PANDA generalizes across different pretrained vision-language models rather than relying on a specific Qwen3-VL backbone.

Table 6: Performance Comparison across Different Multi-modal Large Language Models.

Datasets	ShipReID-2400				VesselReID			
MLLM	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
Qwen3-VL-2B	48.90	54.40	69.10	74.40	75.46	**87.26**	94.98	96.92
LLaVA-OneVision-7B	50.28	56.81	71.53	77.05	71.40	82.61	94.04	96.74
InternVL-8B	**54.46**	**60.62**	**73.80**	**79.32**	**77.33**	86.88	**95.10**	**97.36**

A.3 Effect of Annotator Capacity

Table 7 studies how the capacity of the annotation MLLM affects PANDA. We keep the training pipeline unchanged and vary only the model used to generate privileged attributes. The results show a clear trend: larger annotators produce better final retrieval performance. On ShipReID-2400, using Qwen3-VL-2B for annotation yields 49.58% mAP and 56.78% Rank-1, while Qwen3-VL-8B improves the results to 51.19% mAP and 59.63% Rank-1. A similar trend appears on VesselReID,

Table 7: Performance comparison using different frozen annotator.

Datasets	ShipReID-2400				VesselReID			
Frozen Annotator	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Qwen3-VL-2B	49.58	56.78	71.66	76.62	73.44	85.13	94.04	96.92
Qwen3-VL-4B	50.68	58.92	72.36	76.90	75.79	86.94	94.98	97.05
Qwen3-VL-8B	**51.19**	**59.63**	**72.38**	**77.90**	**76.02**	**88.20**	**95.86**	**97.18**

where Rank-1 improves from 85.13% with Qwen3-VL-2B annotations to 88.20% with Qwen3-VL-8B annotations.

This behavior is consistent with the learning-with-privileged-information perspective of PANDA. The annotation model is used only during training to provide privileged attribute supervision, it is not used at inference. Therefore, a more capable annotator can provide more accurate and discriminative training signals, leading to stronger visual descriptors without increasing deployment cost. The gains are moderate rather than drastic, suggesting that PANDA is not overly dependent on a single annotator, but still benefits from stronger privileged information.

A.4 Interaction with Inference-Time MLLM Reranking

To isolate the effect of the initial retriever from inference-time MLLM reasoning, we replace D2InterNet with PANDA for MUSE and ReID-R while keeping their top-20 reranking protocols unchanged. As shown in Table 8, PANDA provides a much stronger starting point: without reranking, mAP improves from 41.60% to 51.19% on ShipReID-2400 and from 61.70% to 76.02% on VesselReID.

Table 8: Effect of the initial retriever on MLLM-based reranking.

Datasets		ShipReID-2400				VesselReID			
Initial	Reranker	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
D2InterNet	None	41.60	46.00	62.60	70.00	61.70	72.82	89.01	92.91
D2InterNet	MUSE	41.96	46.96	62.93	70.02	63.13	76.69	89.83	93.01
D2InterNet	ReID-R	42.80	48.16	63.93	70.96	64.11	80.44	91.15	92.91
PANDA	None	51.19	59.63	72.38	77.90	76.02	88.20	95.86	97.18
PANDA	MUSE	51.86	**59.77**	72.93	78.08	**76.63**	**89.16**	95.38	97.09
PANDA	ReID-R	**52.03**	59.63	**73.38**	**78.30**	75.96	88.06	**95.89**	**97.20**

On top of PANDA, reranking brings only modest changes. MUSE and ReID-R further improve ShipReID-2400 mAP to 51.86% and 52.03%, respectively. On VesselReID, MUSE slightly improves mAP to 76.63%, while ReID-R remains comparable to PANDA alone. These results show that PANDA's main gain comes from its stronger image-only descriptor, while inference-time MLLM reasoning provides limited additional benefit once the initial retriever is strong.

A.5 Visual Grounding of Hard-Negative Slot Discrimination

As shown in Table 9, we keep the HSD prompt and candidate attributes fixed while changing only the visual input, and define the logit margin as $z_{\text{correct}} - z_{\text{incorrect}}$. After Stage 2, the model shows substantially stronger discrimination when the visual input is consistent with the target, whereas mismatched, negative, or removed visual evidence causes clear performance degradation. Notably, the logit margin reverses sign under the negative-image condition, indicating that the prediction changes according to the visual evidence rather than relying on textual candidate priors alone. The near-chance performance with zero visual input further rules out a strong text-only shortcut. These results directly support the intended role of HSD: it teaches the visual representation to distinguish competing attribute values based on image evidence, thereby strengthening visual grounding for fine-grained attribute discrimination.

A.6 Quality Audit of PAC Annotations

To assess the quality of the privileged attributes produced by PAC, we conduct a full-set automated audit using GPT-5.6-Sol (OpenAI, 2026) as an external evaluator. The evaluator is used only for

Table 9: Visual-grounding control for Hard-Negative Slot Discrimination. Acc. denotes classification accuracy, and Margin denotes the logit margin.

Model Stage	Matched		Mismatched		Counterfactual		Zero Input	
	Acc.	Margin	Acc.	Margin	Acc.	Margin	Acc.	Margin
Before FAS	54.69	0.085	52.34	-0.100	53.13	-0.136	51.56	-0.108
Stage 1 (GAS)	54.69	0.111	53.13	0.075	57.81	0.141	51.56	-0.108
Stage 2 (HSD)	**71.09**	**0.353**	54.84	0.158	39.53	-0.248	51.56	-0.108

post-hoc analysis and is not involved in annotation generation or model training. The complete evaluation prompt is provided in Figure 6 for reproducibility.

PAC Two-View and Counterfactual AI Visual Review Prompt

You are an AI visual reviewer. Evaluate the supplied annotation pair using only the two images. Do not infer hidden details or use external knowledge. **Not visible does not mean absent. This review is not human ground truth.**

Inputs
Two images, an ordered list of attributes with assigned slots, and an optional counterfactual task specifying its target image, positive description, and negative description.

Instructions
1. **Attribute correctness: For each attribute, label each image **S** (supported), **C** (clearly contradicted), or **U** (uncertain or unobservable). Assess the complete attribute; do not split compound descriptions.**
2. **Cross-view consistency: An attribute passes only when both images directly support it.**
3. **Slot assignment: Label the assigned slot **Y** (correct), **N** (incorrect), or **U** (ambiguous):**
- Geometry: vessel type and overall shape.
- Hull: hull colors, paint, and stripes.
- Superstructure: cabins, bridge, doors, and windows.
- Deck: cargo, coverings, hatches, and railings.
- Equipment: cranes, masts, funnels, antennas, lifeboats, and fenders.
- Details: names, numbers, logos, and identifying decorations.
4. **Counterfactual validity: Label the task **V** (valid), **I** (invalid), **U** (uncertain), or **M** (missing). A valid task requires image support for the positive and clear exclusion of the negative for the **same object and property**. Synonymous or compatible descriptions are not valid negatives.**
5. Tag applicable errors: **hallucination, view-specific, wrong slot, too generic, semantically equivalent negative. Multiple tags are allowed. Do not label uncertainty alone as hallucination.**

Output
For each pair, provide:
- Pair ID and a brief visual observation.
- Each attribute's two image labels, slot judgment, and error tags.
- CF verdict and a short justification.
- Notes explaining incorrect or uncertain judgments.

Metrics
Report each dataset–teacher group separately, with numerators and denominators:
- **Attribute precision: At least one S and no C / all attributes.**
- **Two-view consistency: SS / all attributes.**
- **Slot assignment accuracy: Y / all attributes.**
- **Invalid CF rate: I / (V+I+U).**

Report uncertainty and missing CF counts separately. Missing CFs are not valid CFs. Preserve all sampled pairs and do not assume larger models produce better annotations.

Figure 6: Prompt for PAC Annotation Quality Audit.

As shown in Table 10, PAC produces generally reliable attribute annotations across both datasets. Increasing annotator capacity consistently improves attribute precision and slot-assignment accuracy, while cross-view consistency remains strong overall. The quality of counterfactual construction also improves with stronger annotators on ShipReID, whereas invalid counterfactuals are nearly absent on VesselReID.

Table 10: Full-set quality audit of PAC annotations generated by different annotators.

Dataset	Annotator	Attr. Prec. (%) ↑	Cross-view Cons. (%) ↑	Slot Acc. (%) ↑	Invalid CF (%) ↓
ShipReID-2400	Qwen3-VL-2B	79.82	69.88	82.46	19.68
	Qwen3-VL-4B	82.56	79.12	82.56	13.85
	Qwen3-VL-8B	**86.77**	**83.53**	**84.75**	**9.30**
VesselReID	Qwen3-VL-2B	73.81	67.46	86.51	**0.00**
	Qwen3-VL-4B	76.53	75.35	88.61	**0.00**
	Qwen3-VL-8B	**89.63**	**86.47**	**94.49**	0.01

Overall, the results indicate that stronger annotators provide cleaner and more structured privileged supervision, supporting the use of Qwen3-VL-8B for PAC. As the audit is performed automatically, these measurements are primarily used as large-scale quality diagnostics rather than calibrated ground-truth accuracy.

B Qualitative Analysis

B.1 Feature Space Visualization

Figure 7 visualizes the UMAP feature distributions of nine visually similar vessel identities on ShipReID-2400 and VesselReID. Compared with D2InterNet, PANDA produces more compact and better separated identity clusters, indicating that the learned descriptor captures more identity-preserving cues rather than relying only on coarse visual similarity.

The comparison between the two ablated variants further illustrates the complementary roles of Generative Attribute Supervision and Hard-Negative Slot Discrimination. Without Hard-Negative Slot Discrimination, the clusters show stronger overlap between different vessels, suggesting that generative attribute supervision alone is less sufficient for separating highly similar identities. In contrast, without Generative Attribute Supervision, inter-class separation becomes clearer, but samples of the same identity are more scattered. This indicates that hard-negative slot discrimination emphasizes discriminative differences between similar attribute values, while generative supervision helps preserve consistent intra-identity structure.

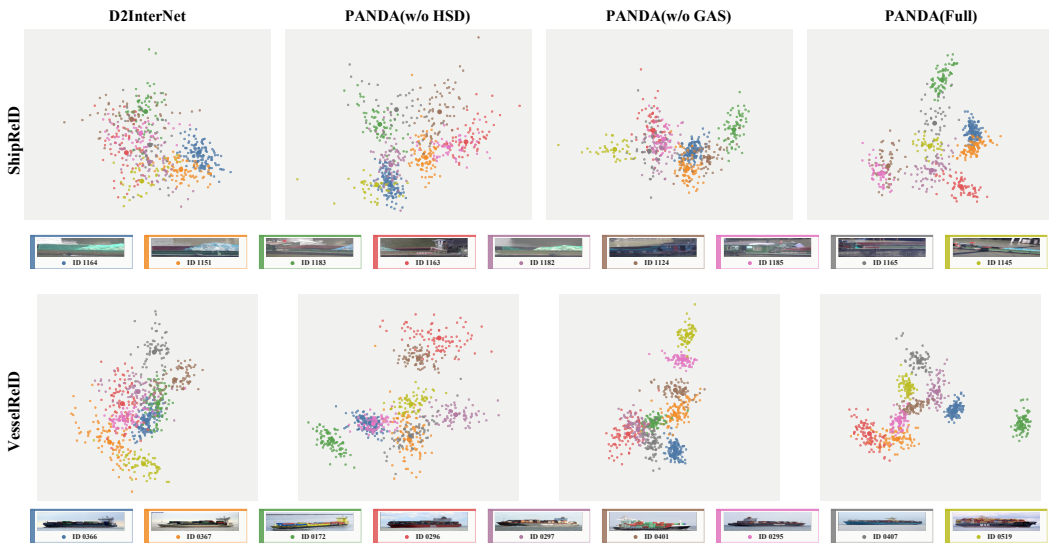


Figure 7: UMAP visualization of feature distributions for nine visually similar vessel identities on ShipReID-2400 and VesselReID. PANDA forms more compact intra-identity clusters and clearer inter-identity separation than D2InterNet and its ablated variants.

The full PANDA model combines both effects. It forms clusters that are both compact within each identity and well separated across identities on both datasets. This visualization is consistent with our design: Generative Attribute Supervision encourages visual tokens to encode cross-view-stable vessel attributes, Hard-Negative Slot Discrimination sharpens sensitivity to fine-grained attribute differences, and Attribute-Aware Fusion retains these cues in the descriptor.

B.2 Retrieval Case Study

Figures 8and 9 show representative retrieval results on ShipReID-2400 and VesselReID, respectively. Green boxes denote correct matches and red boxes denote incorrect matches. We compare D2InterNet, CLIP-ReID, MUSE, ReID-R, and PANDA under the same query.

The results highlight a key limitation of inference-time reranking methods. MUSE reranks the candidates produced by an initial retriever. When the correct match is absent from the top-20 candidate

set, MUSE cannot recover it even with stronger multimodal reasoning. In the ShipReID-2400 example, the correct image appears only when expanding the candidate pool to top-50, showing that reranking is bounded by the recall of the initial retriever.

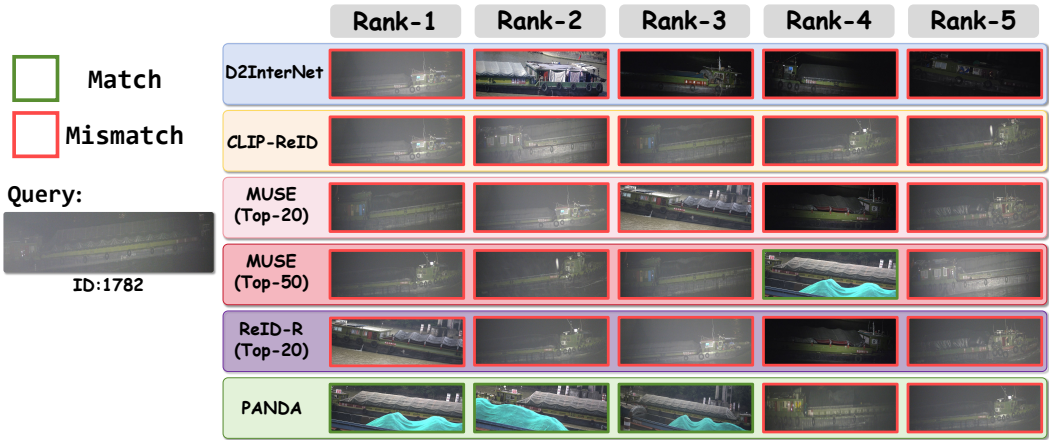


Figure 8: Retrieval case study on ShipReID-2400.

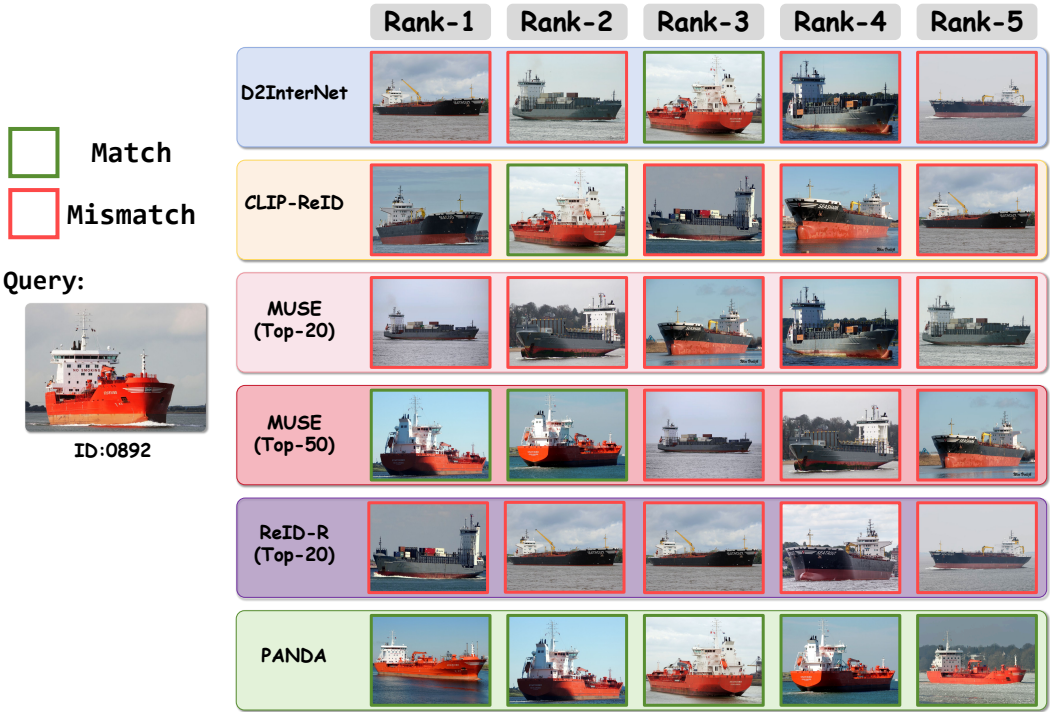


Figure 9: Retrieval case study on VesselReID.

PANDA avoids this dependency because it does not perform inference-time language reranking. Instead, language-derived attributes are used during training to shape the visual descriptor, and retrieval is conducted directly with the image-only descriptor. As shown in both examples, PANDA directly ranks correct matches at the top with its image-only descriptor, whereas the compared visual and reranking methods are misled by globally similar vessels. This supports the central design of PANDA: transferring attribute knowledge into the descriptor during training improves retrieval without relying on candidate-list coverage at inference.

B.3 Attention Visualization

Figure 10 visualizes the attention maps of D2InterNet, PANDA without Hard-Negative Slot Discrimination, PANDA without Generative Attribute Supervision, and the full PANDA. Compared with D2InterNet, PANDA variants attend more consistently to vessel regions that contain identity-related cues, such as hull appearance, superstructure, deck layout, and local equipment.

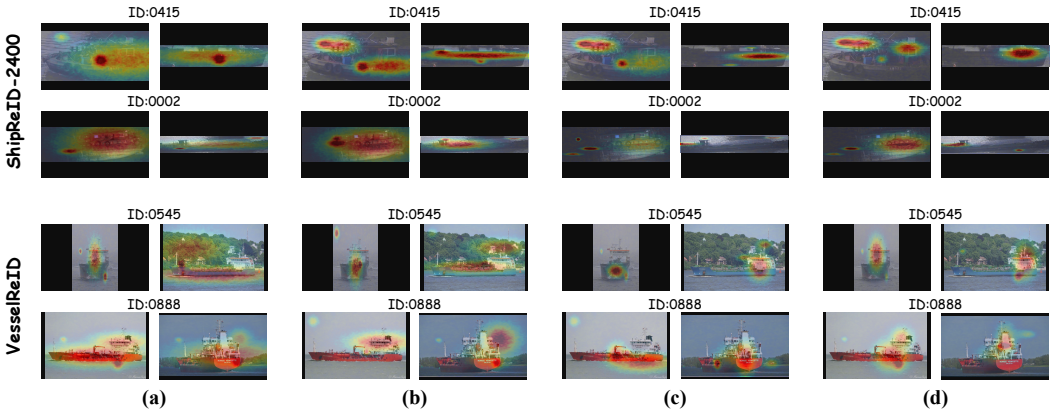


Figure 10: Attention visualization of different methods. (a) D2InterNet. (b) PANDA without Hard-Negative Slot Discrimination. (c) PANDA without Generative Attribute Supervision. (d) Full PANDA.

The two ablated variants reveal different roles of the frozen language-model supervision objectives. Without Hard-Negative Slot Discrimination, the model can still discover discriminative regions, but the activated areas are often broad and less selective. This suggests that generative attribute supervision helps locate stable vessel attributes, but does not sufficiently force the model to focus on fine-grained differences between similar attribute values. Without Generative Attribute Supervision, the attention becomes more localized in some cases but is more likely to shift toward inaccurate or incomplete details, indicating that hard-negative discrimination alone provides contrastive pressure but lacks stable attribute grounding.

The full PANDA model produces more focused and semantically aligned attention maps. It preserves the attribute localization ability provided by Generative Attribute Supervision while using Hard-Negative Slot Discrimination to sharpen the model's sensitivity to fine-grained attribute differences. This qualitative evidence is consistent with our method design: generative supervision teaches the visual encoder what stable vessel attributes to encode, while counterfactual slot discrimination encourages the encoder to attend to the details that distinguish visually similar ships.

C Additional Method Details

C.1 Privileged Attribute Construction

Cross-view attribute generation. PAC uses training images only. For each identity, we sample at most two cross-camera image pairs with a fixed random seed, single-camera identities are excluded. A frozen multimodal annotator extracts vessel attributes shared by both views using the prompt in Figure 11. We parse its JSON response, retrying after Markdown-fence removal and then on the substring between the outermost braces if necessary. Unparseable responses are discarded. Only string entries in `shared_attributes` supply attribute content, neither the overall description nor image-specific observations are added to the slots.

Prompt for Cross-View Attribute Generation

```
You are given two images of the same ship identity from a Ship ReID training set.

Your task is to write a noise-invariant identity description that can supervise each image individually for ship re-identification.

Strict rules:
1. Use only visual cues that are present in BOTH images or can be inferred as persistent structure of the same ship.
2. shared_attributes must contain only common, stable identity cues. Do NOT include anything visible in only one image.
3. If a cue is visible in only one image, put it in image_specific_observations, not in shared_attributes.
4. Ignore viewpoint, crop/partial visibility, weather, lighting, day/night, blur, occlusion, camera angle, water, shore, other ships, and background.
5. Prefer stable cues: hull shape, hull color, waterline color, superstructure position/color, door/window layout, deck layout, railings, masts/cranes/funnels, tire fenders, markings that are clearly present in both images.
6. Include fine-grained discriminative structure when shared by both images: cabin placement, number/arrangement of windows or doors, cargo cover color/shape, deck equipment layout, fender placement, stripe/marking placement, crane/mast/funnel arrangement.
7. Do not describe generic class-only facts unless paired with discriminative detail. For example, "cargo ship" alone is too weak; "flat cargo barge with green tarp and blue cabin near stern" is useful.
8. Avoid uncertain words in shared_attributes such as "possibly", "maybe", "appears", "visible only", "in one image", "in the first image", or "in the second image".

Return compact JSON only:
{
  "identity_description": "2 concise sentences using only common stable identity-related cues, with enough detail to distinguish this ship from similar ships.",
  "shared_attributes": ["common stable cue present in both images", "..."],
  "discriminative_details": ["fine-grained common cue useful for ReID", "..."],
  "image_specific_observations": {
    "image_1_only": ["cue visible only in image 1, if any"],
    "image_2_only": ["cue visible only in image 2, if any"]
  },
  "ignored_noise": ["viewpoint", "lighting", "partial crop", "..."],
  "confidence": "high|medium|low"
}
```

Figure 11: Cross-view prompting for privileged attribute generation.

Attribute filtering. We apply case-insensitive substring filtering with the following exclusion list: *in one, one view, first image, second image, image 1, image 2, other image, only, visible, partially, possibly, maybe, appears, background, shore, weather, lighting, blur, occlusion*. An entire attribute is removed upon any match. A pair is retained if at least three attribute entries survive, irrespective of confidence level or slot occupancy. This is a rule-based acceptance criterion, not a semantic correctness assessment. Each accepted pair produces two image records sharing attributes, confidence metadata, and pair provenance, images occurring in both pairs retain separate records. Figure 12 reports the fraction of training identities covered by the retained annotations, rather than annotation accuracy.

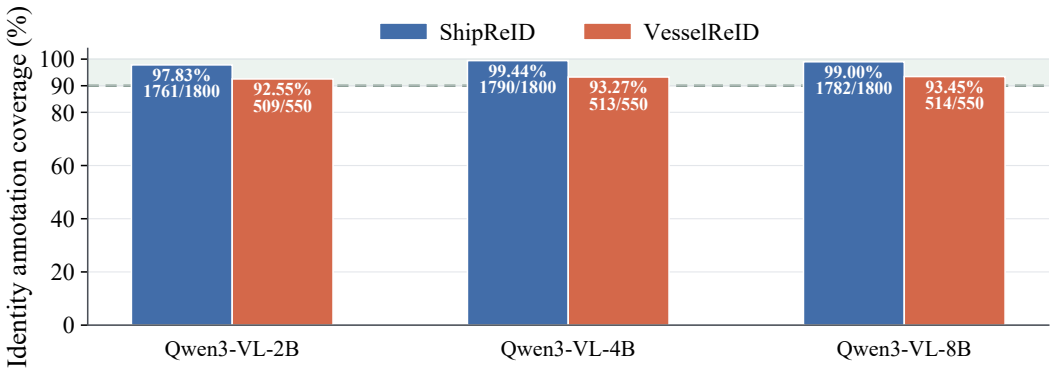


Figure 12: Identity-level PAC coverage with three annotator sizes. Bar labels show coverage and annotated/total training identities.

Slot encoding. Retained attributes are trimmed and processed in their original order. Let $\mathcal{M}(a)$ contain the slots among the first five in Table 11 whose keywords match attribute a. Matching is case-insensitive, literal, and bounded by regular-expression word boundaries, internal phrase spacing

is unchanged. The entire attribute is assigned to

$$g(a) = \min\big(\mathcal{M}(a) \cup \{6\}\big). \tag{12}$$

Thus, the first matching slot takes precedence, and slot 6 is the fallback. Attribute capitalization is preserved, without lemmatization, synonym expansion, compound-sentence splitting, or guaranteed deduplication. All six slots are stored, with empty lists for unfilled slots. Non-empty slots are serialized in priority order as "slot name: attributes", using spaces in place of underscores, semicolons between attributes, and periods between slots, followed by a final period. Accepted inputs use the retained attribute list and do not invoke the constructor's fallback to the overall description.

Table 11: Complete slot-assignment rules. Priority follows the row order, unmatched attributes enter the fallback slot.

	Slot	Keywords
1	Vessel type and geometry	barge, carrier, tanker, boat, vessel type, hull shape, flat-bottomed, long
2	Hull appearance	hull, waterline, stripe, paint
3	Superstructure	superstructure, cabin, bridge, wheelhouse, door, window
4	Deck layout	deck, cargo, hatch, tarp, cover, container
5	Equipment	crane, mast, funnel, antenna, lifeboat, fender, tire
6	Distinctive details	All unmatched attributes

Single-slot counterfactuals. For identity p, donor identities $q \neq p$ are ranked by the Jaccard similarity of their retained attribute sets:

$$J(p, q) = \frac{|A_p \cap A_q|}{|A_p \cup A_q|}. \tag{13}$$

We select a target slot from the non-empty slots and replace its content with that of the corresponding donor slot, leaving all other slots unchanged. Replacements with identical or subset token sets are rejected. The resulting original and counterfactual slot texts, (c_i^+, c_i^-), provide the HSD candidates.

C.2 Frozen-LM Attribute Supervision

Generative Attribute Supervision. The frozen language model receives prompt tokens, visual tokens, and the serialized attribute description in this order. Only answer tokens contribute to the generative loss, prompt, visual, and padding positions are masked. Figure 13 gives the supervision prompts.

Prompt Templates for Frozen-LM Attribute Supervision

Prompt Template for Generative Attribute Supervision

Describe the vessel using only stable identity-related visual attributes such as hull color, superstructure, deck layout, cargo cover, cabin, doors, windows, fenders, cranes, and vessel type. Ignore lighting, weather, water, background, camera viewpoint, image quality, and crop.

Prompt Template for Hard-Negative Slot Discrimination

Choose the stable vessel attribute value that exactly matches the image. Ignore viewpoint, lighting, crop, water, and background.
Target stable identity attribute: *<slot name>*. Compare only the candidate values for this attribute.
Image visual tokens appear before the candidates.
Candidate A: *<candidate A>*
Candidate B: *<candidate B>*
Answer with A or B only.

Figure 13: Prompts for generative attribute supervision and hard-negative slot discrimination.

Hard-Negative Slot Discrimination. HSD asks which candidate slot matches the image. The positive and counterfactual candidates are randomly assigned to options A and B, with each of the two possible orders sampled with probability 0.5 and the target label adjusted accordingly. The per-record loss ℓ_i^{CF} is computed from the option-token logits with the correct option as the target, no answer text is supplied as input. Its contribution is weighted by the PAC sample weight and the frequency of its counterfactual target slot.

Sample weighting. PAC stores a heuristic weight γ_i for each image record. With $N_{\mathrm{attr},i}$ retained attribute entries and $c_i = \min(1, N_{\mathrm{attr},i}/6)$, we compute

$$\begin{aligned} \widehat{\gamma}_i &= b_i(0.65 + 0.35c_i)u_i n_i, \\ \gamma_i &= \mathrm{round}_4\big(\max\{0.25, \widehat{\gamma}_i\}\big), \end{aligned} \tag{14}$$

where rounding retains four decimal places. The confidence factor b_i equals 1.00, 0.75, or 0.45 for high, medium, or low confidence, respectively. An absent confidence key uses 0.75, explicit null or unrecognized values use 0.65. The provenance factor u_i is 1 for a non-empty source-pair identifier and 0.85 otherwise. The noise factor n_i is 0.75 if the joined attributes match any of *weather, water, background, viewpoint, camera, lighting, night, daytime, sky, image quality, crop*, and 1 otherwise, using case-insensitive word-boundary matching. Here $N_{\mathrm{attr},i}$ counts attributes, not occupied slots, and provenance indicates source availability rather than cross-view agreement. These factors define a training heuristic, not a calibrated probability of annotation correctness.

Slot-frequency weighting and reduction. Let n_s count CF records targeting slot s during line-by-line indexing of the annotation file, and let $n_{\max} = \max_s n_s$. For an observed target slot s_i, we use

$$\beta_s = \sqrt{\frac{n_{\max}}{n_s}}, \qquad w_i = \max(0, \gamma_i \beta_{s_i}). \tag{15}$$

The counts include repeated image records and are neither mini-batch frequencies nor positive-slot occupancy counts. For the valid CF set $\mathcal{B}_{\mathrm{CF}}$ in a mini-batch, define $M = |\mathcal{B}_{\mathrm{CF}}|$ and $W = \sum_{i \in \mathcal{B}_{\mathrm{CF}}} w_i$. HSD minimizes

$$\mathcal{L}_{\mathrm{HSD}} = \begin{cases} \dfrac{\sum_{i \in \mathcal{B}_{\mathrm{CF}}} w_i \ell_i^{\mathrm{CF}}}{W}, & W > 0, \\ \dfrac{1}{M} \displaystyle\sum_{i \in \mathcal{B}_{\mathrm{CF}}} \ell_i^{\mathrm{CF}}, & M > 0,\ W = 0, \\ 0, & M = 0. \end{cases} \tag{16}$$

Rejected targets must be excluded from $\mathcal{B}_{\mathrm{CF}}$, since zero total weight triggers an unweighted mean rather than a zero loss. This normalization adjusts relative within-batch contributions: a shared slot-frequency factor cancels in a single-slot batch, so the weighting does not ensure balanced gradients across batches.

Optimization. Stage 1 trains the visual LoRA with generative supervision. Stage 2 initializes from Stage 1 and applies weighted HSD with a low-weight generative replay term. Only the visual LoRA are updated, the language model, token embeddings, and output head remain frozen throughout.

C.3 Attribute-Aware Fusion

AAF maps the adapted visual tokens to an image-only retrieval descriptor through query extraction, entropy-modulated interaction, and descriptor construction.

Query extraction. Six learnable semantic queries and two identity queries attend to normalized, projected visual tokens. A residual feed-forward block then refines the query representations:

$$\begin{aligned} V &= W_{\text{in}}\text{LN}_{\text{token}}(T), \\ Q_0 &= \text{LN}_{\text{init}}([Q_s, Q_u]), \\ R_q &= \text{MHA}(\text{LN}_q(Q_0), V, V), \\ Q_1 &= Q_0 + \text{Dropout}(R_q), \\ Q_2 &= Q_1 + \text{Dropout}\big(\text{FFN}(\text{LN}_f(Q_1))\big), \\ S &= Q_2[1{:}6], \qquad U = Q_2[7{:}8]. \end{aligned} \tag{17}$$

where S and U denote the semantic and identity tokens, respectively, and $[\,,\,]$ denotes concatenation.

Entropy-modulated interaction. As illustrated in Figure 14, semantic tokens provide a gated residual correction to the identity anchor. Writing $n(\cdot)$ for ℓ_2 normalization, we compute

$$\begin{aligned} a_0 &= \text{BN}_0\left(W_0\, n\left(\frac{[n(U_1), n(U_2)]}{\sqrt{2}}\right)\right), \\ s'_k &= W_S\text{LN}_S(S_k), \\ u'_r &= W_U\text{LN}_U(U_r), \\ A_{rk} &= \text{softmax}_k\left(\frac{(u'_r)^\top s'_k}{\sqrt{128}}\right), \\ m_r &= \sum_{k=1}^{6} A_{rk} s'_k. \end{aligned} \tag{18}$$

Normalized attention entropy and a learned gate modulate the correction:

$$\begin{aligned} H_r &= -\frac{\sum_{k=1}^{6} A_{rk}\log A_{rk}}{\log 6}, \\ q_r &= 0.25 + 0.75(1 - H_r), \\ \rho_r &= \sigma(W_{\text{gate}}[u'_r, m_r] + b_{\text{gate}}), \\ \xi_r &= \rho_r q_r\big(\tanh(u'_r) \odot \tanh(m_r)\big), \\ \widetilde{a} &= a_0 + W_{\text{res}}\text{GELU}([\xi_1, \xi_2]). \end{aligned} \tag{19}$$

Zero-initializing W_{res} makes the initial output equal to the uncorrected anchor.

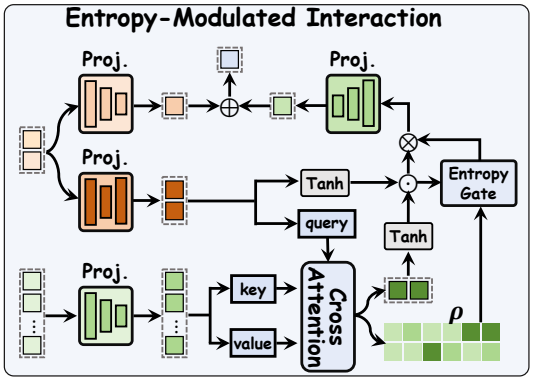


Figure 14: Entropy-modulated interaction between semantic and identity tokens.

Descriptor construction. The identity branch combines attention-pooled global evidence, its compact projection, and the corrected anchor:

$$\begin{aligned} \alpha_n &= \mathrm{softmax}_n(w_g^\top T_n), \\ g &= \sum_n \alpha_n T_n, \qquad p = W_p g, \\ a &= \mathrm{BN}_a\left(W_a \begin{bmatrix} \sqrt{0.70}\, n(\widetilde{g}) \\ \sqrt{0.15}\, n(\widetilde{p}) \\ \sqrt{0.15}\, n(\widetilde{a}) \end{bmatrix}\right). \end{aligned} \tag{20}$$

The semantic branch retains six semantic query vectors in separate coordinates. The final descriptor is

$$\begin{aligned} v_k &= n(\mathrm{LN}_{\mathrm{slot}}(W_{\mathrm{slot}} n(S_k))), \\ s &= [v_1, \ldots, v_6]/\sqrt{6}, \\ f &= \begin{bmatrix} \sqrt{0.60}\, n(a) \\ \sqrt{0.40}\, n(s) \end{bmatrix} \in \mathbb{R}^{1792}, \end{aligned} \tag{21}$$

where $a \in \mathbb{R}^{1024}$, $v_k \in \mathbb{R}^{128}$, and $s \in \mathbb{R}^{768}$. Branch energy weights are fixed, and f is used directly for nearest-neighbor retrieval.

Algorithm 1 Training pipeline of PANDA

Require: Training data $\mathcal{D}_{\mathrm{train}}$; frozen annotator $\mathcal{M}_{\mathrm{ann}}$ and LM p_ϕ.
Ensure: Visual retriever $x \mapsto f(x)$.

Privileged Attribute Construction
for each training identity **do**
  Sample cross-camera pairs and extract shared attributes with $\mathcal{M}_{\mathrm{ann}}$.
  Clean, assign slots, and serialize attributes into Y_i.
  Construct (c_i^+, c_i^-) by single-slot replacement using equation 13.
end for
Stages 1–2: GAS → HSD with replay
Initialize E_θ from the pretrained vision tower; freeze base weights.
for $\mathcal{L} \in (\mathcal{L}_{\mathrm{GAS}}, \mathcal{L}_{\mathrm{stage2}})$ in order **do**
  for each training minibatch **do**
    Encode $T_i = E_\theta(x_i)$ and evaluate $\mathcal{L}$ through frozen p_ϕ.
    Update only visual adapters using equation 3 or equation 5.
  end for
end for
Stage 3: Attribute-Aware Fusion
Remove p_ϕ and attach AAF.
for each training minibatch **do**
  Encode $T_i = E_\theta(x_i)$.
  Compute $f_i = \mathrm{AAF}(T_i)$ using equation 17–equation 21.
  Update visual adapters and AAF with $\mathcal{L}_{\mathrm{ReID}}$ in equation 10.
end for
return E_θ and the AAF retrieval head.

C.4 Training Pipeline

Algorithm 1 summarizes training: PAC constructs descriptions and counterfactuals offline, frozen-LM supervision trains the visual adapters, and AAF trains the final image-only descriptor after removing the language model. At inference, neither language model, prompts, nor attribute annotations are required. A single visual forward pass and AAF produce the 1792-dimensional descriptor for nearest-neighbor ranking.

D Detailed Experimental Setup

D.1 Datasets

ShipReID-2400. ShipReID-2400 is a large-scale ship re-identification benchmark collected from a real-world intelligent waterway monitoring system. It contains 17,241 images from 2,400 ship

identities, captured over 53 months under different camera views, locations, times, and weather conditions. Following the official split, 1,800 identities are used for training, 300 for validation, and 300 for testing. The official split contains 12,988 training images, 691 validation query images, 706 test query images, and 4,253 gallery images.

VesselReID. VesselReID contains 30,587 images from 1,248 vessel identities, captured across different times, places, viewpoints, and weather conditions. We follow the standard evaluation protocol for vessel re-identification and report retrieval performance on the corresponding test split.

D.2 Evaluation Metrics

We report standard retrieval metrics, including mAP, Rank-1, Rank-5, and Rank-10 accuracy. mAP measures the overall ranking quality, while Rank-k reports whether at least one correct match appears in the top-k retrieved results. For efficiency analysis, we report GFLOPs and latency. Latency is measured as the inference time required to process one image on an NVIDIA Pro6000 GPU.

D.3 Comparison Baselines

We compare PANDA with three groups of baselines.

Generic visual ReID methods. This group includes representative methods that improve image-only ReID descriptors through metric learning, stronger backbones, part-aware modeling, or object-level feature interaction. DPM (Hermans et al., 2017) uses triplet-loss based metric learning as a strong descriptor baseline. TransReID (He et al., 2021) introduces transformer-based object re-identification. DC-Former (Li et al., 2023b) learns diverse and compact transformer representations. AdaSP (Zhou et al., 2023) improves ReID with adaptive sparse pairwise supervision, while MSINet (Gu et al., 2023) models multi-scale interactions for object ReID. We also include PFD-Net (Wang et al., 2022), a pose-guided feature disentangling method. These baselines evaluate whether PANDA improves over strong visual descriptor learning.

Vehicle and ship-oriented ReID methods. Since ship ReID shares challenges with vehicle ReID, such as viewpoint variation, local structural cues, and cross-camera appearance changes, we include several vehicle- and ship-oriented methods. FED (Qiao et al., 2020) is an early marine vessel re-identification method based on global-local feature fusion. UFDN (Qian et al., 2022) learns decoupled vehicle features without predefined part annotations, and HRCN (Zhao et al., 2021) is introduced with the VERI-Wild vehicle ReID benchmark for large-scale vehicle retrieval. D2InterNet (Liu et al., 2025) is a recent ship-specific baseline proposed with ShipReID-2400, designed to handle scale variation and discriminative local features in ship images. These methods provide domain-relevant comparisons for vessel and ship retrieval.

Language-enhanced and MLLM-based ReID methods. We further compare with methods that incorporate language or large vision-language models into ReID. CLIP-ReID (Li et al., 2023a) exploits CLIP-style vision-language knowledge for image ReID without requiring concrete text labels. Instruct-ReID (He et al., 2024) formulates ReID as an instruction-conditioned retrieval task, while LVLM-ReID (Wang et al., 2026) explores large vision-language models for person re-identification. MUSE (Li & Gong, 2026) and ReID-R (Zhang et al., 2026) perform inference-time multimodal reasoning using Qwen3-VL-2B (Bai et al., 2025a) and Qwen2.5-VL-7B (Bai et al., 2025b), respectively. In our evaluation, both use D2InterNet (Liu et al., 2025) as the initial retriever and apply their MLLM-based reranking or reasoning modules to the top-20 retrieved candidates.

D.4 Implementation Details

We instantiate PANDA with the vision tower of Qwen3-VL-4B-Instruct (Bai et al., 2025a). The base visual encoder is frozen, and LoRA adapters are inserted into the visual backbone with rank 16, scaling factor 32, and dropout 0.05. We use batch size 8 with 4 images per identity in all training stages. All experiments are conducted on a single server equipped with two NVIDIA RTX PRO 6000 Blackwell GPUs.

Privileged attributes are constructed only from the training split. For each training identity, we sample at most two cross-camera image pairs with a fixed seed. A frozen Qwen3-VL-8B-Instruct (Bai et al., 2025a) annotator generates shared vessel attributes, which are cleaned, assigned to six slots, and serialized into positive descriptions. Single-slot counterfactuals are constructed by replacing one non-empty slot with the corresponding slot from an attribute-similar donor identity.

PANDA is trained in three stages. In the first stage, visual adapters are trained for one epoch with generative attribute supervision. In the second stage, training continues for one epoch with hard-negative slot discrimination and a low-weight generative replay term of 0.15. The language model, language embeddings, and language-model head remain frozen in both stages. We use AdamW with weight decay 10^{-4} and base learning rates 1×10^{-4} and 5×10^{-5} for the first and second stages, respectively.

In the third stage, the language model is removed and Attribute-Aware Fusion is trained for 60 epochs with standard ReID losses. The trainable components include visual LoRA adapters, query tokens, pooling layers, projection layers, fusion layers, and classification heads, while the base visual encoder remains frozen. The final descriptor has 1,792 dimensions, consisting of a 1,024-dimensional identity branch and a 768-dimensional attribute branch, with energy weights 0.60 and 0.40. During inference, all language components and privileged attributes are discarded, and retrieval is performed by nearest-neighbor search over the final visual descriptor.

D.5 Direct Attribute Supervision

To isolate the effect of the frozen-LM supervision interface, we construct a controlled baseline using the same PAC annotations and visual backbone as PANDA. Given visual tokens T_i, we apply mean pooling followed by lightweight slot-wise prediction heads to directly predict the PAC attributes with a multi-label binary cross-entropy loss:

$$h_i = \mathrm{LN}\left(\frac{1}{N}\sum_{n=1}^{N} T_{i,n}\right), \qquad \mathcal{L}_{\mathrm{attr,i}} = \frac{1}{|\mathcal{K}_i|}\sum_{k\in\mathcal{K}_i} \mathrm{BCE}(W_k h_i + b_k, y_i^{(k)}), \tag{22}$$

where $\mathcal{K}_i$ denotes the non-empty PAC slots. The baseline follows the same two-epoch supervision schedule as FAS, while updating only the visual LoRA adapters and temporary attribute heads. The attribute heads are discarded before Stage 3, after which the same AAF module and ReID training objective are used. Thus, both settings share the PAC supervision, visual backbone, subsequent ReID training, and inference architecture, and differ only in the training-time interface used to transfer attribute information to the visual encoder.

D.6 Holistic Description SFT

Holistic Description SFT uses the same frozen-LM generative interface as GAS but replaces the slot-structured PAC target with the `identity_description` generated during privileged attribute construction. The description provides a holistic natural-language summary of stable identity-related vessel cues. Only description tokens contribute to the autoregressive loss, the language model remains frozen and gradients update the visual adapters. The language model is discarded before Stage 3, after which the same AAF and ReID training procedure is applied. This baseline isolates the effect of holistic language-generation supervision from PANDA's structured attribute supervision.