Thermal-Det: Language-Guided Cross-Modal Distillation for Open-Vocabulary Thermal Object Detection

Supplementary Material

S1. Implementation Details

We build our framework on Grounding DINO, extending it with Thermal Adapters and TTAH for thermal domain adaptation. All experiments use MMDetection v3.2 with mixed precision and gradient checkpointing. A pretrained Grounding DINO (Swin-T) serves as the frozen RGB teacher, while the thermal student is initialized from the same weights and trained on (1) synthetic thermal data from F-ViTA and (2) paired RGB–thermal datasets (MMPD, Multi-Spectral Stereo, M^3FD). Training uses roughly 250K paired thermal images. Input images are resized to 800×800. The detector processes `p4` and `p5` encoder features (resized to 27×27 and 20×20) concatenated into one token sequence. The CLIP text encoder is frozen, while the adapters, detection head, and TTAH are optimized jointly with detection, distillation, and alignment losses. Training runs on $8\times$ RTX A6000 GPUs with a total batch size of 16 for 150K iterations. We use AdamW with a 2×10^{-4} learning rate, 0.05 weight decay, cosine scheduling, a 2K warm-up, and mixed KD:Synthetic batches at a 3:2 ratio. All baseline methods are evaluated using their official pretrained weights under a shared zero-shot thermal inference protocol, with no fine-tuning applied; applying the synthetic pretraining or distillation pipeline to these baselines would alter their methods and conflate comparisons.

Table S1. Zero-shot detection results on thermal datasets SMOD, MFAD, and LLVIP.

Method	SMOD		MFAD		LLVIP	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
G-DINO	0.135	0.226	0.093	0.169	0.522	0.746
MM-GDINO	0.140	0.237	0.099	0.177	0.561	0.800
LLMDet	0.133	0.229	0.102	0.182	0.517	0.745
Ours	**0.152**	**0.267**	**0.100**	**0.180**	**0.566**	**0.856**
G-DINO (FT)	0.274	0.357	0.194	0.354	0.706	0.979

S2. Additional detection results

In this section, we provide extended quantitative results to complement those presented in the main paper. We first report additional zero-shot detection results on the SMOD, MFAD, and LLVIP datasets, which were not included in the main manuscript. As shown in Table S1, our method consistently outperforms existing RGB open-vocabulary detectors across all three benchmarks and evaluation metrics, demonstrating its robustness and generalization ability in diverse thermal domains.

Tab. S1 and Tab. S2 also include the fully fine-tuned results for G-DINO (FT) across all thermal datasets evaluated in this work, covering FLIR-Aligned, FLIR-V2, CAMEL, Utokyo, SMOD, MFAD, and LLVIP. For completeness, we list alongside them the corresponding zero-shot performance of the RGB-trained open-vocabulary models evaluated in the main paper, specifically G-DINO, MM-GDINO, and LLMDet. Zero-shot results for GLIP, T-Rex2, and YOLO-World were already presented in the main manuscript and are therefore omitted here.

Comparing the G-DINO (FT) results with the zero-shot baselines highlights the level of improvement achievable when supervised thermal annotations are available. Across all datasets, fine-tuned G-DINO achieves significantly higher performance, reflecting the benefits of domain-specific training. At the same time, the difference between zero-shot G-DINO and G-DINO (FT) provides a useful reference for understanding the difficulty of transferring RGB models to the thermal domain without supervision.

Our method is designed to reduce this gap using synthetic thermal supervision and RGB to thermal cross-modal distillation rather than manual labeling. The combined comparisons across all seven datasets present a comprehensive picture of both zero-shot and fully fine-tuned performance and show how closely our approach approaches the supervised upper bound while remaining annotation-free.

S3. Related works

Open-vocabulary object detection. Open-vocabulary object detection (OVD) enables recognition beyond fixed categories through large-scale vision–language pre-training. Early works such as GLIP [22] and Grounding DINO [26] aligned visual regions with text to achieve open-set detection. Recent methods, including YOLO-World [3], T-Rex2 [18], and LLMDet [6], improved efficiency and semantic grounding using one-stage architectures, prompt synergy, and large language model supervision. However, these advances remain confined to the RGB domain, and extending them to thermal imagery demands new strategies for cross-modal alignment and domain adaptation.

Synthetic thermal data and domain adaptation. Thermal object detection suffers from limited labeled data, motivating synthetic generation and domain adaptation to transfer knowledge from RGB imagery. Early studies such as SSTN [31] and Domain-Adaptive Pedestrian Detection [10] used self-supervised and adversarial methods to reduce modality gaps, while ThermalSynth [28] and Ham-

Table S2. Zero-shot detection results on thermal datasets FLIR-Aligned, FLIR-V2, CAMEL, and Utokyo.

Method	FLIR-Aligned			FLIR-V2			CAMEL			Utokyo		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
GLIP	0.251	0.471	0.226	0.025	0.041	0.028	0.186	0.324	0.242	0.049	0.093	0.046
T-Rex2	0.276	0.514	0.255	0.033	0.048	0.036	0.213	0.345	0.274	0.049	0.101	0.043
YOLO-World	0.266	0.487	0.241	0.029	0.044	0.032	0.197	0.336	0.256	0.050	0.101	0.043
G-DINO	0.337	0.636	0.313	0.081	0.144	0.078	0.482	0.729	**0.543**	0.050	0.093	0.169
MM-GDINO	0.354	0.619	**0.371**	0.042	0.060	0.048	0.273	0.369	0.340	0.047	0.095	0.042
LLMDet	0.359	0.628	0.348	0.048	0.075	0.051	0.383	0.560	0.439	0.050	0.102	0.045
Ours	**0.372**	**0.664**	0.359	**0.096**	**0.173**	**0.091**	**0.511**	**0.758**	0.525	**0.065**	**0.137**	**0.054**
G-DINO (FT)	0.414	0.749	0.414	0.200	0.345	0.203	0.478	0.684	0.533	0.129	0.252	0.115

rell and Karlholm [11] improved realism via simulation and GAN-based translation. Meta-UDA [39] introduced meta-learning for unsupervised adaptation across environments. More recent foundation-guided models, including F-ViTA [34] and ThermalGen [41], achieve semantically consistent visible-to-thermal synthesis, marking a shift toward scalable cross-modal generation. Yet, most approaches treat thermal adaptation and semantic grounding separately, and our work unifies them through synthetic supervision, language alignment, and cross-modal distillation for zero-shot thermal detection. Together, advances in open-vocabulary detection and synthetic thermal generation lay the groundwork for scalable perception beyond labeled data. Building on these trends, our work unifies language supervision, synthetic thermal pre-training, and cross-modal alignment to achieve zero-shot thermal object detection.

S4. Pre-training datasets and backbones

To contextualize the capabilities of the baseline open-vocabulary detectors evaluated in this work, Table S3 summarizes their backbone architectures and pre-training corpora. Most RGB open-vocabulary models, including GLIP, Grounding-DINO, MM-GDINO, and LLMDet, rely on Swin-T backbones trained on large-scale grounding datasets such as Objects365, GoldG, and various captioning or region-level corpora. YOLO-World instead adopts a YOLOv8-L backbone combined with mixture-of-caption datasets to enhance efficiency and prompt alignment, while T-Rex2 is trained on a diverse 10M-image collection aggregated from multiple sources. These pre-training differences influence the semantic coverage, grounding quality, and transferability of each model, and highlight that all baselines are optimized primarily for RGB imagery. This underscores the challenge of applying them directly to thermal data and motivates our approach for bridging the modality gap through synthetic thermal supervision and cross-modal distillation.

Table S3. Backbone architectures and pre-training datasets used by baseline open-vocabulary detectors evaluated in this work.

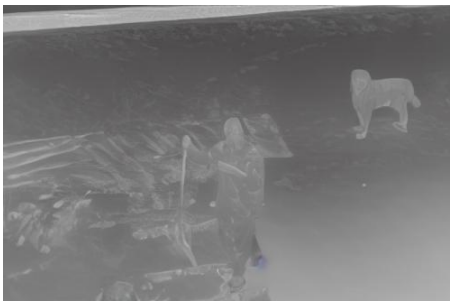
Method	Backbone	Pre-training data
GLIP [22]	Swin	O365, GoldG, Cap4M
T-Rex2 [18]	Swin	10M data from various resources
YOLO-World [3]	YOLO	O365, GoldG, CC3M
G-DINO [26]	Swin	O365, GoldG, Cap4M
MM-GDINO	Swin	O365, GoldG, GRIT, V3Det
LLMDet [6]	Swin	GroundingCap-1M
Ours	Swin	Synthetic GroundingCap-1M M^3FD, MMPD, MS^2

S5. TTAH Drift Regularization Sensitivity

We analyze the sensitivity of TTAH to the drift regularization weight λ_{drift} by varying it over the range $\{0, 0.25, 1, 5, 10\}$ on FLIR-Aligned. Aside from λ_{drift}, all other loss terms are normalized to comparable value ranges. The results show stable behavior across this range: the model peaks at $\lambda_{\text{drift}} = 0.25$ (AP=0.259) and degrades gradually for larger values (AP=0.247 at $\lambda_{\text{drift}} = 1$, AP=0.226 at $\lambda_{\text{drift}} = 5$), indicating robustness rather than sensitivity to this hyperparameter. Setting $\lambda_{\text{drift}} = 0$ removes the drift constraint entirely, allowing unconstrained adaptation that hurts generalization. These runs were conducted on a single GPU due to resource constraints.

S6. Synthetic Data Domain Shift Analysis

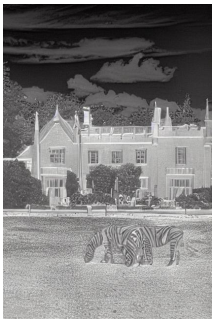
To characterize the domain shift between our synthetic thermal data and real thermal benchmarks, we compute FID scores between the synthetic dataset generated via F-ViTA and the FLIR-Aligned validation set. The FID between our synthetic dataset and FLIR-Aligned is 41.40, comparable to the FID between two real thermal benchmarks, FLIR-V2 and FLIR-Aligned (38.74), indicating a similar domain shift magnitude. Caption adaptation removes RGB-specific descriptors only—such as color terms and lighting conditions—while preserving object identity and modality-invariant spatial relations.



A person stands in a snowy outdoor area, holding a cold, shovel-shaped object whose metal blade appears significantly cooler than the surroundings. The individual's warm head and torso contrast with the insulated, cooler areas of their clothing. Beside them, a dog displays a distinct warm body signature with a cooler collar visible around its neck. The ground shows cold snow with slightly warmer linear patterns indicating tire tracks, while the background includes a cold, snow-covered paved surface and snow-covered vegetation on the left.

Two warm human figures walk across a cool, grassy hillside, their bodies showing strong heat signatures from the head and torso with noticeably cooler patterns across their layered clothing. The individual on the left carries a cooler, bag-shaped object whose low thermal emission contrasts with the warmer hand holding it. The terrain beneath them appears mostly cool, with mixed patches of slightly warmer vegetation and scattered cooler leaf material. In the background, multiple trees appear as large, mostly cool vertical structures, including evergreens with very low heat emission and deciduous trees showing faint, uneven warmth. The overall scene indicates a cool outdoor environment with soft thermal contrast across the landscape.

A warm cat sits inside a cool, bowl-shaped sink basin, its body showing a concentrated heat signature with the head and torso emitting the strongest thermal contrast. The sink surrounding the cat appears uniformly cool, with the metal faucet registering as even colder. Spread around the countertop are numerous toiletry items displaying varying thermal profiles—most of them cool, with a few showing mild warmth likely due to recent handling or sun exposure. The countertop itself emits a broad, stable cool temperature, while the mirror behind it reflects minimal thermal information. The cat's warm body sharply contrasts the cool surfaces around it, making it the dominant thermal feature of the scene.



Two large, warm zebra-shaped heat signatures stand on a cool grassy field, their bodies emitting consistent warmth with slightly cooler patterns along the legs and extremities. The ground beneath them appears broadly cool with subtle variations caused by grass density and soil composition. Behind the animals, a large stone mansion rises as a predominantly cool structure, its walls and roof retaining very little heat and displaying uniform low-temperature surfaces. The chimneys and architectural details show the same cold thermal characteristics. Above the building, the sky registers as a broad, cold background with no significant thermal emission. The warm zebras create a striking contrast against the cool grass and the colder stone architecture behind them.

Figure S1. Qualitative thermal examples with descriptions.

Functionally, a detector trained exclusively on synthetic data (without any subsequent distillation on real paired data) achieves 0.372 mAP on FLIR-Aligned, with near parity on the *person* class relative to fully fine-tuned G-DINO. This confirms that the synthetic initialization provides meaningful semantic coverage and serves as an effective starting point for cross-modal distillation, with remaining mismatches corrected through subsequent training on real RGB–thermal pairs.

S7. Limitations

Our approach has several limitations. First, the synthetic thermal data used during pre-training does not fully reproduce sensor characteristics such as emissivity variations, temperature gradients, and noise patterns, leaving a residual domain gap between synthetic and real thermal imagery. Second, the cross-modal distillation stage depends on the accuracy of the RGB teacher detector, and any errors in the teacher's predictions can be transferred to the thermal student and limit zero-shot performance. A further limitation arises from the textual supervision used during pre-training. Because the synthetic dataset is constructed from region-level grounding phrases, the model receives many more region-level annotations than detailed image-level descriptions. This imbalance can cause the language-conditioned components to generate shorter or less informative captions, even when detailed descriptions are requested. Incorporating more descriptive region-level text or stronger image-level supervision could help address this limitation and further improve semantic grounding.

S7.1. Visualizations of synthetic training data

In Fig. S1, we illustrate several examples of the synthetic thermal images and their corresponding textual annotations used during pre-training. These samples are drawn from our synthetic dataset, where thermal images are produced through RGB-to-thermal translation and paired with region-level grounding phrases derived from the original RGB captions. The annotations accurately identify the main objects within each scene, demonstrating that the grounding-based text supervision provides reliable category signals for open-vocabulary thermal detection.