Completing Missing Modalities: Synthetic Data for RGB–Infrared–Thermal–Text Person Re-Identification

Muhammad Umair Muhammad Hammad Musaddiq Zhou Jun Ahmad Muhammad
University of Electronic Science and Technology of China, Chengdu, China
{muhammadumair894, hammadmusaddiq, ahmadjameel7171}@gmail.com, zhouj@uestc.edu.cn

Abstract

Multi-modal person re-identification is increasingly constrained by a data bottleneck rather than by backbone design alone. In realistic deployments, robust retrieval may require aligned visible RGB, infrared (IR), thermal infrared (TI), and textual descriptions, yet acquiring such data directly with synchronized multi-sensor capture is expensive, logistically difficult, and often infeasible when some sensor streams are unavailable. We address this problem from a synthetic-data-centric perspective and present MM-ReID, a unified dataset built through controlled modality completion. MM-ReID contains 8,537 identities, 855,022 RGB/IR/TI images, and 832,966 captions. Instead of synthesizing new identities, we preserve the original identity space and generate only missing modalities. This strategy yields balanced four-modality supervision at a scale that would be difficult to obtain from direct multi-sensor acquisition alone. Overall, 46.1% of the visual data are original sensor observations and 53.9% are generated cross-modal views, enabling large-scale aligned training without requiring a fully instrumented capture pipeline for every identity. We further use this dataset to examine whether a model can learn robust modality-agnostic representations from mixed real and synthetic supervision. Using Cross-Modal Semantic Anchoring (CMSA), a ViT equipped with frozen vision-language anchors and caption-aware objectives, we show that synthetic modality completion provides effective supervision for unified RGB-IR-TI-text learning, reaching 79.4 mAP, 82.7 R1 on Market1501-MM. These results indicate that carefully curated synthetic data can compensate for missing sensor streams and reduce data-collection burden. https://github.com/muhammadumair894/MM-ReID

1. Introduction

Person re-identification (ReID) aims to match the same individual across non-overlapping cameras under varying viewpoints, illumination, and appearance conditions. Early benchmarks such as CUHK03 and Market-1501 established the standard RGB-only setting, while MSMT17 introduced larger environmental variation and more realistic camera deployments [9, 17, 19]. Subsequent datasets addressed specific challenges individually, including clothing changes [14, 16], natural-language-based retrieval [7, 20], and cross-spectrum matching across visible, near-infrared, and thermal imagery [18]. Meanwhile, recent models such as CLIP-ReID, Instruct-ReID, AIO, and ReID5o have pushed ReID toward stronger vision–language and multi-modal alignment [3, 6, 8, 21]. Despite these advances, the data ecosystem remains fragmented: language supervision is typically separated from infrared and thermal sensing, multi-spectral datasets remain limited in scale and diversity, and there is still no widely used benchmark where RGB, IR, thermal infrared (TI), and text are jointly available under diverse environments and conditions.

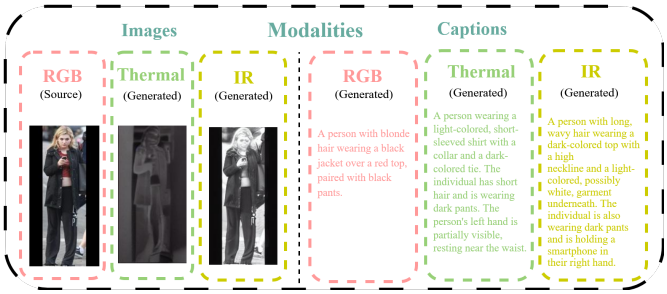

Figure 1. Example of generated modalities (TI, IR) from source RGB and captions for RGB, TI, and IR images.

This limitation is primarily a data-centric challenge. Constructing a large-scale four-modality benchmark through direct acquisition requires multiple sensors, repeated capture across environments and times of day, identity-preserving matching across cameras, and extensive manual verification. Prior work on visible–infrared ReID already highlights that collecting such data is costly, labor-intensive, and often restricted by privacy considerations [1]. More broadly, synthetic-data research consis-

tently identifies annotation cost, capture difficulty, and privacy constraints as key bottlenecks for large-scale visual datasets [2, 4]. These challenges are even more pronounced in person-centric tasks, where identity consistency must be preserved while capturing diverse appearance variations.

Synthetic data provides a promising alternative when used to complement scarce real observations rather than replace them entirely. Prior studies show that carefully designed synthetic supervision can reduce annotation effort and still transfer effectively to real-world tasks [2, 4]. In multimodal scenarios, synthetic pre-training has improved performance in applications such as egocentric interaction detection when real data are limited [5]. Diffusion models have significantly lowered the barrier to generating high-quality synthetic data. Parameter-efficient diffusion pipelines can now produce useful training samples on modest hardware [13], and text-guided generation pipelines enable identity-preserving person variations for ReID training [1, 11]. These developments suggest that the key question is no longer whether synthetic data can assist multi-modal learning, but how it can be used systematically to complete missing sensing modalities.

In this work, we adopt this perspective and formulate synthetic data generation as a *modality-completion* problem for person ReID. Instead of synthesizing new identities, we preserve the identity space of existing ReID datasets while generating the missing modalities required to align RGB, IR, TI, and text. Specifically, infrared and thermal views are synthesized from RGB images using F-ViTA [12], while identity-relevant captions are generated using Describe Anything [10]. This process produces MM-ReID, a unified dataset containing 8,537 identities, 855,022 RGB/IR/TI images, and 832,966 captions. The dataset intentionally mixes real and synthetic data: real sensor observations anchor each identity, while generated modalities increase cross-modal coverage where direct acquisition is unavailable.

Text plays an important role in this unified setting. Such semantic cues can remain informative even when color or texture information is degraded in infrared or thermal imagery. Incorporating language therefore both enables text-based queries and encourages the visual representation to organize identities around semantically meaningful attributes rather than modality-specific appearance.

To evaluate whether synthetic modality completion supports unified multi-modal learning, we employ Cross-Modal Semantic Anchoring (CMSA), a ViT-based framework that uses frozen vision–language embeddings as semantic anchors and introduces caption-aware losses to separate visually ambiguous identities.

Our contributions are threefold.

- We present a synthetic-data-centric formulation of multi-modal person ReID and construct MM-ReID, a unified benchmark spanning RGB, infrared, thermal infrared, and text at a scale difficult to achieve through direct multi-sensor capture.
- We demonstrate how foundation-model-based generation can perform identity-preserving modality completion, synthesizing IR/TI views from RGB and generating captions to provide semantic supervision as shown in Figure 1.
- We show that a single model trained on this mixed real–synthetic benchmark can learn robust modality-agnostic representations across RGB, IR, TI, and text.

2. Related Work

2.1. Person Re-Identification Benchmarks

Person re-identification has traditionally been developed on RGB-only benchmarks such as CUHK03 and Market-1501, which established standard evaluation protocols under controlled settings [9, 19]. MSMT17 later increased realism by introducing larger scale, more cameras, and stronger illumination variation [17]. However, RGB-only datasets fail to capture challenges common in real deployments, including clothing changes, low-light sensing, and cross-modal retrieval.

To address appearance variation, clothing-change datasets such as LTCC and PRCC emphasize long-term identity consistency under outfit changes [14, 16]. Language-based retrieval benchmarks such as CUHK-PEDES and UFineBench introduced natural-language descriptions as complementary supervision, enabling retrieval using semantic attributes rather than pixels alone [7, 20]. Multispectral datasets further extended ReID toward infrared and thermal sensing. RGBNT201 and related benchmarks provide aligned visible, infrared, and thermal observations, but remain limited in scale and diversity [18**?**]. Overall, existing benchmarks typically address only one challenge at a time RGB scale, clothing change, language, or cross-spectrum sensing rather than jointly supporting RGB, IR, thermal, and text.

2.2. Synthetic Data for Vision and Multimodal Learning

Synthetic data has emerged as an important alternative when real-world data collection and annotation are expensive or restricted. Early work demonstrated that fully synthetic training can significantly reduce annotation cost while offering strong control over dataset statistics, although bridging the synth-to-real gap remains challenging [4]. Large-scale synthetic benchmarks further showed that diversity and task-relevant variability are crucial for transferring synthetic supervision to real-world tasks [2].

Recent diffusion-based approaches have improved the practicality of synthetic generation. For example, FLORA

demonstrates that parameter-efficient diffusion tuning can generate useful synthetic datasets with much smaller computational budgets [13]. Synthetic data has also proven beneficial in multimodal scenarios. Leonardi *et al.* showed that synthetic multimodal pre-training can improve performance on real-world industrial tasks when in-domain data are limited [5]. These results suggest that synthetic data is most effective when used to complement real observations and fill gaps in sensing or annotation.

2.3. Synthetic Data for Person Analysis and Cross-Modal ReID

Within person-centric vision, synthetic data has been explored to address appearance changes, missing modalities, and limited sensor coverage. For instance, recent work on clothes-changing ReID generates harder identity-preserving training samples to increase appearance diversity [11]. In visible–infrared ReID, DiVE demonstrates that RGB–IR synthesis can improve cross-modal matching by decoupling identity information from modality-specific appearance [1].

2.4. Unified Multi-Modal ReID

Recent ReID models increasingly incorporate vision–language priors and unified multimodal learning. CLIP-ReID demonstrated that pretrained vision–language models can improve person representation learning even without manual text annotations [8]. Instruct-ReID further generalized the problem formulation through instruction-driven retrieval [3]. More recent frameworks such as AIO and ReID5o aim to support multiple sensing modalities within a single architecture rather than using separate pipelines for each modality pair [6, 21].

Despite these advances, the progress of multimodal models has outpaced the availability of suitable datasets. Language-based benchmarks typically lack infrared or thermal imagery, while multispectral datasets rarely include detailed textual supervision. Consequently, the primary bottleneck remains data-centric rather than algorithmic.

3. Methodology

Our methodology is designed around a data-centric premise: in multi-modal person re-identification, the main bottleneck is often not the backbone itself, but the lack of aligned observations across sensing modalities. Collecting synchronized RGB, infrared (IR), thermal infrared (TI), and language descriptions for the same identity at scale is expensive and operationally difficult. Following recent evidence that synthetic data is most effective when used to complement scarce real data rather than replace it entirely [2, 4, 5], we formulate dataset construction as an *identity-preserving modality completion* problem.

Our pipeline has two components. First, we aggregate fragmented public ReID datasets and complete missing modalities using off-the-shelf foundation models, producing aligned RGB-IR-TI-text tuples without altering identity labels. Second, we validate the usefulness of the resulting mixed real-synthetic benchmark using a single unified model, Cross-Modal Semantic Anchoring (CMSA), which learns modality-agnostic and language-aware embeddings from RGB, IR, TI, and text jointly. This separation is deliberate: the primary contribution of this workshop paper is the synthetic-data pipeline, while CMSA serves as a downstream test of whether the generated modalities provide useful supervision.

3.1. Problem Formulation

Let $\mathcal{D}_0 = \{(x_i^R, y_i)\}_{i=1}^N$ denote an initial collection of person images with identity labels, where x_i^R is an RGB observation and y_i is the identity. Some source datasets also contain real infrared or thermal observations, while others contain only RGB imagery [9, 16**?** –19]. Our goal is to construct a unified multi-modal dataset

$$\mathcal{D} = \{(x_i^R, \tilde{x}_i^I, \tilde{x}_i^T, \tilde{t}_i, y_i)\}_{i=1}^N, \tag{1}$$

where $\tilde{x}_i^I$ and $\tilde{x}_i^T$ denote infrared and thermal views that may be either real or synthesized, and $\tilde{t}_i$ denotes a natural-language description that is automatically generated. Importantly, we do *not* synthesize new identities. Instead, we preserve the original identity space and only complete missing modalities for identities already present in the source benchmarks.

This formulation is attractive for data-centric multi-modal learning for two reasons. First, it avoids the need for a fully instrumented capture setup for every subject. Second, it converts fragmented uni-modal and partially multi-modal datasets into aligned supervision suitable for training a single model across sensing conditions.

3.2. Identity-Preserving Modality Completion

Source benchmark aggregation. We begin from publicly available person ReID datasets that collectively cover conventional RGB ReID, clothing variation, and multi-spectral sensing [9, 16**?** –19]. This aggregation strategy gives us broad identity diversity, multiple environments, and heterogeneous sensing conditions, but it also yields incomplete modality coverage: some datasets contain only RGB images, whereas others contain RGB plus one or more additional spectra.

Infrared and thermal synthesis. For each RGB image lacking a paired infrared or thermal observation, we generate the missing modalities using F-ViTA [12]. Formally, given an RGB image x_i^R, we obtain

$$\hat{x}_i^I = G_I(x_i^R), \qquad \hat{x}_i^T = G_T(x_i^R), \tag{2}$$

where $G_I(\cdot)$ and $G_T(\cdot)$ denote the translation functions used to produce IR and TI views. When real IR/TI images are available in the source data, we retain those sensor observations and do not replace them with generated images. In this way, synthetic views act as *modality densification* rather than as a substitute for all real sensing.

Automatic caption generation. To add a fourth modality, we generate natural-language descriptions using Describe Anything [10]. For each image or aligned visual tuple, we produce a short identity-relevant description

$$\hat{t}_i = C(x_i^R, \tilde{x}_i^I, \tilde{x}_i^T), \tag{3}$$

where $C(\cdot)$ denotes the captioning model. In practice, we guide caption generation toward identity-discriminative attributes such as clothing type, color and pattern, carried accessories, and coarse appearance cues, while suppressing irrelevant background details. This is important because our goal is not general scene captioning; instead, we require language that is useful for person retrieval.

Human verification and filtering. Because synthetic supervision is only useful when it preserves identity-relevant content, we insert a human-in-the-loop quality control stage after generation. Generated IR/TI images are inspected for translation failures such as hallucinated edges, missing body parts, or severe blur. Likewise, captions are reviewed through prompt refinement so that they emphasize person attributes rather than scene clutter. This step reduces obvious synthetic artefacts before the data are used for downstream training. The identity preservation in our setting refers to preserving the ReID label and identity-discriminative body-level cues, not to generating photorealistic or face-faithful reconstructions. Since many ReID datasets contain low-resolution pedestrian crops where faces are not reliable identity cues, our filtering prioritizes body silhouette, clothing layout, carried accessories, and pose consistency.

MM-ReID Dataset Applying this pipeline yields MM-ReID (as shown in Table 1), a unified dataset containing 8,537 identities, 855,022 RGB/IR/TI images, and 832,966 captions. Of the visual data, 394,487 images (46.1%) are original sensor observations and 460,535 images (53.9%) are generated cross-modal views, while all captions are automatically generated. Crucially, the identity space is kept fixed throughout: synthetic images and captions are attached only to existing identities rather than to newly hallucinated subjects. This allows the dataset to expand modality coverage without corrupting label semantics.

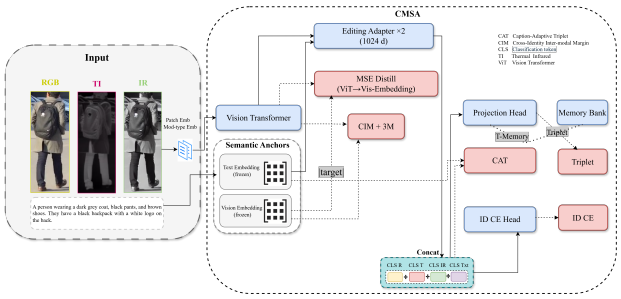

Figure 2. Proposed Cross-Modal Semantic Anchoring (CMSA) model

3.3. Text as a First-Class Modality

Text is introduced not as auxiliary metadata, but as a genuine retrieval modality and a semantic regularizer. Language-based person retrieval benchmarks have shown that short descriptions can encode identity-relevant information that is difficult to preserve consistently in pixels alone [7, 20]. In our setting, this is even more important because RGB, IR, and TI do not preserve the same visual cues equally well. Color may be distinctive in RGB but disappear in infrared imagery; thermal shape and body contour may remain stable while texture is reduced. Text helps bridge these inconsistencies by expressing higher-level semantic attributes that can remain informative across sensing domains. To reduce caption redundancy, we constrain the caption prompt to focus on identity-relevant attributes, including clothing type, dominant colors, visible accessories, body pose, and coarse appearance cues. We also remove captions that are dominated by background descriptions or generic templates. Although this does not provide a full linguistic diversity analysis, it reduces the risk that the text modality collapses into repetitive or scene-biased supervision.

3.4. Cross-Modal Semantic Anchoring

To test whether the completed dataset supports unified multi-modal learning, we adopt Cross-Modal Semantic Anchoring (CMSA) as the downstream validation model as shown in Figure 2. CMSA is intentionally lightweight: instead of training separate heads or modality-specific pipelines, it uses a single ViT backbone together with frozen vision-language anchors.

Let a mini-batch be

$$\mathcal{B} = \{(x_i^R, x_i^I, x_i^T, t_i, y_i)\}_{i=1}^{B}. \tag{4}$$

The RGB, IR, and TI images are patch-embedded and passed to a student Vision Transformer. Modality-type tokens are added so that the backbone can distinguish the three visual streams while still processing them within a unified architecture. In parallel, frozen foundation encoders

Dataset	RGB Img	Thermal Img	IR Img	Total Img	RGB Cap	Thermal Cap	IR Cap	Total Cap	#ID	Category	Environment
Celeb-ReID	34 186	34 186 ✓	34 186 ✓	102 558	34 186 ✓	34 186 ✓	34 186 ✓	102 558	1 052	Clothing Change	Indoor
MSMT	126 441	125 806 ✓	126 405 ✓	378 652	126 441 ✓	125 806 ✓	126 405 ✓	378 652	4 101	Traditional	Indoor + Outdoor
Market1501	22 971	22 186 ✓	22 186 ✓	67 343	22 234 ✓	22 186 ✓	22 186 ✓	66 606	1 501	Traditional	Outdoor
Cuhk03	14 096	14 092 ✓	14 092 ✓	42 280	14 096 ✓	14 092 ✓	14 092 ✓	42 280	1 360	Traditional	Indoor
PRCC	33 698	33 698 ✓	33 698 ✓	101 094	33 698 ✓	33 698 ✓	33 698 ✓	101 094	221	Clothing Change	Indoor
Real2	21 320	21 319	0	42 639	21 320 ✓	0	0	21 320	101	Real-world Complex	Outdoor
RGBNT201	4 116	4 116	4 116	12 348	4 116 ✓	4 116 ✓	4 116 ✓	12 348	201	Multi-Modal	Outdoor (Day/Night)
Market1501-MM	36 036	36 036	36 036	108 108	36 036 ✓	36 036 ✓	36 036 ✓	108 108	–	Multi-Modal	Outdoor
Total	**292 864**	**291 439**	**270 719**	**855 022**	**292 127**	**270 120**	**270 719**	**832 966**	**8 537**	–	–

Table 1. Entries annotated with Green ✓denote synthetic samples generated by F-ViTA (for TI/IR images) or Describe Anything (for captions); unmarked entries correspond to original images reused from the source datasets. The Market1501-MM ID count is omitted due to duplication with Market1501.

provide visual and textual anchor representations. The student ViT produces modality-specific visual tokens, while the frozen text encoder yields caption embeddings that serve as semantic anchors.

The core intuition of CMSA is simple: the student should not learn three separate modality-dependent spaces. Instead, it should learn a shared embedding geometry in which RGB, IR, TI, and text representations of the same identity are semantically compatible. To achieve this, CMSA concatenates the modality-specific student tokens with the frozen language anchor and refines them through lightweight adapter MLPs before projection to the final embedding space. The student backbone and adapters are trainable, while the teacher encoders remain frozen.

Formally, for sample i, let $v_i \in \mathbb{R}^{d_v}$ denote the student visual embedding and $c_i \in \mathbb{R}^{d_t}$ denote the frozen caption embedding. These embeddings are then optimized with standard identification and metric learning losses together with two caption-aware objectives, described below.

3.5. Caption-Aware Optimization

Caption-Adaptive Triplet (CAT). The first objective addresses visually subtle but semantically overlapping negatives. For samples i and j, we compute caption similarity as

$$s_{ij} = \frac{c_i^\top c_j}{\|c_i\|_2 \|c_j\|_2}. \tag{5}$$

Let $p(i)$ and $n(i)$ denote the hardest positive and hardest negative for anchor i in the visual embedding space, and define the cosine distance

$$d_{ij} = 1 - \cos(v_i, v_j). \tag{6}$$

Instead of using a fixed triplet margin, CAT makes the margin proportional to caption overlap:

$$m_i^{\text{CAT}} = \alpha \, s_{i\,n(i)}. \tag{7}$$

The resulting loss is

$$\mathcal{L}_{\text{CAT}} = \frac{1}{B} \sum_{i=1}^{B} \Big[d_{i\,p(i)} - d_{i\,n(i)} + m_i^{\text{CAT}} \Big]_+. \tag{8}$$

Thus, if two identities have similar captions, the model is required to separate them by a larger visual margin. CAT is especially useful in cross-modal retrieval because visually ambiguous identities often remain ambiguous after spectrum translation.

Caption-aware CIM-T. The second objective explicitly enlarges inter-identity margins across modalities when caption similarity is high. For an anchor-negative pair (a, n) with $y_a \neq y_n$, we first define the text-conditioned margin

$$m_{an}^{\text{CIM-T}} = m_0 + \beta s_{an}, \tag{9}$$

where m_0 is the base margin and β controls how strongly textual overlap increases the required separation. Let $f_a^{z_1}$ and $f_n^{z_2}$ denote modality-specific features for modalities $z_1, z_2 \in \{R, I, T\}$. We measure inter-modal discrepancy by

$$D_{z_1,z_2}(a, n) = \|f_a^{z_1} - f_n^{z_2}\|_1. \tag{10}$$

CIM-T then penalizes the worst cross-modal pair:

$$\Delta_{an} = \max_{z_1,z_2 \in \{R,I,T\}} \left| m_{an}^{\text{CIM-T}} - D_{z_1,z_2}(a, n) \right|, \tag{11}$$

and the batch loss is

$$\mathcal{L}_{\text{CIM-T}} = \frac{1}{|\mathcal{P}|} \sum_{(a,n) \in \mathcal{P}} \Delta_{an}, \tag{12}$$

where $\mathcal{P}$ is the set of anchor-negative pairs in the mini-batch. In effect, CIM-T prevents semantically similar but identity-mismatched samples from collapsing together across RGB, IR, and TI.

4. Experiments and Results

Unlike conventional ReID evaluations that treat the dataset as fixed, our experiments are designed to answer a data-centric question: *can synthetic modality completion provide supervision that transfers to real multi-sensor person re-identification?* This distinction is important. In synthetic-data research, the decisive test is not whether a model performs well on generated data alone, but whether mixed real-synthetic training improves robustness on real target domains [1, 2, 4, 5]. We therefore evaluate MM-ReID not

Table 2. Performance of CMSA trained on MM-ReID.

Dataset	mAP	R-1	R-5	R-10	± std
Market1501	74.1	82.7	96.2	97.5	0.033
MSMT17	45.4	70.0	82.6	86.5	0.031
CUHK03	70.3	72.2	87.0	92.4	0.025
PRCC	35.3	34.5	39.9	43.0	0.017
RGBNT201	61.2	64.0	75.8	81.6	0.037
Market1501-MM	79.4	82.7	97.5	98.8	0.022

only as a large unified dataset, but also as a vehicle for studying whether missing infrared, thermal, and language modalities can be completed synthetically without compromising downstream generalization.

4.1. Experimental Setup

We train on the MM-ReID dataset described in Table 1, which unifies RGB, IR, TI, and text. This design follows the same principle emphasized in recent synthetic-data work: generation is most useful when it complements scarce real observations rather than replacing them entirely [1, 2, 5].

We report results on six public datasets spanning RGB-only, clothing-change, and real multi-sensor settings: Market1501, MSMT17, CUHK03, PRCC, RGBNT201, and Market1501-MM [9, 16–19]. Following standard ReID practice, performance is measured using mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-1, Rank-5, and Rank-10.

Our CMSA model has a ViT-Base backbone. Input RGB, IR, and TI images are resized to 256×128, and we apply standard augmentation with random horizontal flipping and random erasing. Training lasts 90 epochs with AdamW, batch size 128, initial learning rate 3×10^{-5}, and weight decay 0.05. Each mini-batch contains four images per identity. The overall objective combines ID loss, triplet loss, Caption-Adaptive Triplet (CAT), Caption-aware CIM-T, and an ℓ_2 distillation term.

4.2. Unified Learning from Mixed Real-Synthetic Data

Table 2, reports the main quantitative results. A single CMSA model trained on MM-ReID achieves strong performance across all six benchmarks, including both conventional RGB datasets and real multi-sensor datasets. In particular, CMSA reaches 61.2 mAP / 64.0 Rank-1 on RGBNT201 and 79.4 mAP / 82.7 Rank-1 on Market1501-MM, which are the most relevant benchmarks for assessing synthetic-to-real transfer in our setting. Std denotes the standard deviation of Rank-1 performance across repeated runs. [15].

These results support two observations. First, synthetic modality completion does not confine the model to a narrow generated domain; instead, the learned representation transfers across diverse conditions, including day/night sensing, clothing variation, and cross-camera changes. Second, the benefit of synthetic data is not limited to enlarging the dataset numerically. Because each identity remains tied to real observations, the generated IR/TI views act as *cross-modal densification*, allowing the model to observe richer modality correspondences for the same person [1].

Table 3. Comparison with prior methods on Market1501-MM. Best results are highlighted in bold.

Method	mAP	R-1	R-5	R-10
Single-modal				
MLFN	42.7	68.1	87.1	92.0
HACNN	42.9	69.1	86.6	92.2
OSNet	39.7	69.3	86.7	91.3
Multi-modal				
HAMNet	60.0	82.8	92.5	95.0
PFNet	60.9	83.6	92.8	95.5
IEEE	64.3	83.9	93.0	95.7
HTT	67.2	81.5	95.8	97.8
EDITOR	77.4	**90.8**	96.8	98.3
TIENet	67.4	86.1	94.1	96.0
DeMo	78.0	90.7	96.8	98.2
CMSA (ours)	**79.4**	82.7	**97.5**	**98.8**

4.3. Comparison with SOTA

Table 3 compares CMSA against both single-modal and recent multi-modal methods on Market1501-MM. Earlier multi-modal methods raise performance into the low-to-mid 60s, while more recent 2024-2025 methods push performance into the high-70 range. CMSA achieves the best mAP (79.4) and the best top-k retrieval consistency (97.5 Rank-5 and 98.8 Rank-10), although EDITOR and DeMo remain stronger on Rank-1.

This comparison is important for the synthetic-data argument. If the generated modalities were merely producing visually plausible but semantically weak supervision, one would not expect the resulting model to outperform specialized real-data multi-modal systems on a challenging benchmark. Instead, CMSA achieves the strongest overall ranking quality in terms of mAP and top-k retrieval, suggesting that synthetic modality completion improves the *consistency of the gallery ranking*, not just the first match.

4.4. Efficiency and Scalability

A practical advantage of the proposed setting is that the enriched benchmark can still be leveraged by a single reasonably efficient model. Table 4 reports the computational cost of three CMSA variants on Market1501-MM. The trainable

Table 4. Computational complexity of CMSA on Market1501-MM.

Backbone	Trainable Params (M)	FLOPs (G)	Latency (ms)	Memory (MB)
ViT-Base	97.75	58.67	4.58	2169.91
ViT-Small	59.93	43.75	3.41	2009.14
DeiT-Small	33.25	33.21	2.59	1897.04

student backbone ranges from 33.25M to 97.75M parameters, while the total forward cost ranges from 33.21 to 58.67 GFLOPs for one sample consisting of three visual modalities and one caption. On an NVIDIA Tesla V100 32GB GPU, latency ranges from 2.59 ms to 4.58 ms per sample, with a memory footprint below 2.2 GB in all cases.

4.5. Discussion

Taken together, the experiments support a clear data-centric conclusion: synthetic modality completion is a practical way to alleviate the scarcity of aligned RGB, IR, TI, and text data in person re-identification. The main gain does not come from inventing additional identities, but from increasing the number of aligned modality observations per identity. This richer cross-modal density improves retrieval on real multi-sensor benchmarks and enables one unified model to operate across heterogeneous sensing conditions.

At the same time, our results should not be interpreted as evidence that synthetic data fully replaces real sensing. The present study shows that synthetic IR/TI views and generated captions can *bridge missing modalities* effectively, but a more controlled causal analysis, for example, comparing real-only, mixed real-synthetic, and synthetic-only training under the same protocol, would further clarify how much of the gain comes from modality completion, additional scale, and semantic caption supervision, respectively. We leave that fine-grained disentanglement to future work.

5. Data and Code Availability

Data and code are available from the authors upon reasonable request, subject to source-dataset licensing and benchmark curation constraints.

References

[1] Zhen Dai, Xiaolong Wang, and Liang Zheng. Dive: Disentangled visible-infrared person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2, 3, 5, 6

[2] Marcello Fabbri, Simone Calderara, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5, 6

[3] Xiangyu He, Zhiwei Luo, and Jian Zhang. Instruct-reid: Instruction-driven person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3

[4] Stefan Hinterstoisser, Vincent Lepetit, and Slobodan Ilic. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2019. 2, 3, 5

[5] Francesco Leonardi, Marco Bertini, and Alberto Del Bimbo. Exploiting multimodal synthetic data for egocentric human-object interaction detection in industrial scenarios. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 2, 3, 5, 6

[6] He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. *arXiv preprint arXiv:2405.04741*, 2024. 1, 3

[7] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4

[8] Shuang Li, Dongxu Li, and Jingdong Wang. Clip-reid: Exploiting vision-language model for image-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1, 3

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2, 3, 6

[10] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*, 2025. 2, 4

[11] Hankun Liu, Yujian Zhao, and Guanglin Niu. Try harder: Hard sample generation and learning for clothes-changing person re-id. *arXiv preprint arXiv:2507.11119*, 2025. 2, 3

[12] Jay N. Paranjape, Celso de Melo, and Vishal M. Patel. F-vita: Foundation model guided visible to thermal translation. *arXiv preprint arXiv:2504.02801*, 2025. 2, 3

[13] Alvaro Patricio, Atabak Dehban, and Rodrigo Ventura. Flora: Efficient synthetic data generation for object detection in low-data regimes via finetuning flux lora. *ArXiv*, abs/2508.21712, 2025. 2, 3

[14] Xiangyu Qian, Yan Huang, Ning Xu, and Yi Yang. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 1, 2

[15] Muhammad Umair, Zhou Jun, Muhammad Hammad Musaddiq, and Ahmad Muhammad. Unifying rgb, thermal infrared, infrared, and text for person re-identification: A multi-modal dataset and vision-language transformer. *Information Processing and Management*, 63(1):104589, 2026. 6

[16] Fang Wan, Ping Hu, and Shaogang Gong. Person re-identification by contour sketch under clothing change. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 6

[17] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[18] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3529–3537, 2021. 1, 2

[19] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 3, 6

[20] Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22010–22019, 2024. 1, 2, 4

[21] Jialong Zuo, Yongtai Deng, Mengdan Tan, Rui Jin, Dongyue Wu, Nong Sang, Liang Pan, and Changxin Gao. ReID5o: Achieving omni multi-modal person re-identification in a single model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. 1, 3