

UAV-CC: A Fine-Grained Change Captioning Benchmark for UAV Remote Sensing

Anonymous Authors

Abstract—Remote sensing change captioning (RSCC) generates natural language descriptions of semantic differences between bi-temporal image pairs, enabling human-readable monitoring of land-cover evolution. All existing RSCC benchmarks, including LEVIR-CC [1], CCEXpert [2], and UniRS [3], are built on satellite imagery at ground sampling distances (GSD) of approximately 0.5 m. At this resolution, only scene-level changes—building construction, road expansion, vegetation clearing—are spatially discriminable, and captions accordingly describe coarse-grained events.

UAV platforms achieve GSDs of 5–30 cm, revealing individual vehicles, construction equipment, temporary structures, and other objects that are sub-pixel at satellite scale. This resolution difference is not merely quantitative: it enables a qualitatively new category of change description unavailable from satellite imagery. Yet no benchmark exists to measure, train, or compare models at this level of granularity.

We introduce UAV-CC, the first UAV-resolution change captioning benchmark. UAV-CC comprises 2,077 bi-temporal UAV image pairs sourced from genuine multi-date aerial surveys, annotated with 10,385 object-level change captions generated by GPT-4o and verified by human annotators. Using UAV-CC, we conduct systematic experiments across six baselines spanning zero-shot large vision-language models (LVMs), satellite-trained models applied cross-scale, and super-resolved satellite models. Our key findings are: (1) zero-shot satellite-trained models achieve only 27.1 CIDEr on UAV-CC, a 64.6-point gap below our UAV-specific fine-tuned model at 91.7 CIDEr; (2) super-resolution of satellite imagery closes this gap by only 39.2 CIDEr points, leaving a persistent domain deficit; and (3) the proposed Object-Mention Recall (OMR) metric correlates more strongly with human preference (Spearman $\rho = 0.73$) than CIDEr ($\rho = 0.54$) on object-level descriptions.

UAV-CC and all code are publicly available at <https://github.com/anonymous/uavcc>.

Index Terms—UAV remote sensing, change captioning, vision-language model, benchmark, object-level description

I. INTRODUCTION

Change captioning—generating natural language descriptions of differences between two images of the same scene at different times—has emerged as a useful interface between automated remote sensing analysis and human decision-makers. Unlike binary change detection maps, captions convey semantics: what changed, where, and by how much. Applications span construction progress monitoring, disaster assessment, urban planning audit, and facility inspection.

Existing remote sensing change captioning (RSCC) research has developed entirely within the satellite domain.

UAV-CC: Change Captioning at UAV vs Satellite Resolution

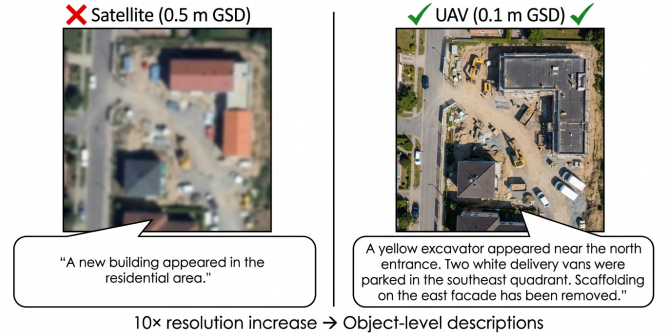


Fig. 1: Resolution determines description granularity. At satellite scale (0.5 m GSD, left), change captions describe coarse scene-level events. At UAV scale (0.1 m GSD, right), the same scene reveals object-level changes invisible from above. UAV-CC is the first benchmark for the latter.

The dominant benchmark, LEVIR-CC [1], provides 10,077 bi-temporal pairs at 0.5 m GSD covering building changes in Texas, USA. Subsequent benchmarks—DUBAI-CCD [4], CCEXpert [2], ChangeIMTI [5]—and models—UniRS [3], BTCChat [6], GeoLLaVA [7]—all operate at this scale. At 0.5 m GSD, individual vehicles, construction equipment, or parked containers occupy at most a few pixels, preventing any object-specific captioning.

UAVs occupy a distinct sensing niche. Modern commercial drones routinely achieve 5–30 cm GSD, bringing individual objects into sharp focus: a forklift in a specific bay, scaffolding added to the east facade, a delivery truck occupying a previously empty parking space. These object-level observations are precisely the type of actionable change information demanded by construction managers, emergency responders, and security analysts—but current RSCC models cannot produce them, and no benchmark can measure whether a model does so correctly.

This paper makes three contributions.

- 1) **UAV-CC benchmark.** We introduce the first UAV-resolution bi-temporal change captioning benchmark: **2,077** image pairs at 5–30 cm GSD, **10,385** object-level captions, with **25%** of pairs human-verified. Pairs are sourced from UAV-BCD [8] and the YZDS dataset [9], providing authentic multi-date UAV surveys with genuine semantic changes.
- 2) **Systematic resolution analysis.** We show that satellite-

trained models and super-resolution preprocessing cannot substitute for UAV-native training data. The residual performance gap after super-resolution and cross-domain fine-tuning is **25.4** CIDEr points, pointing to a fundamental distributional mismatch beyond spatial frequency content.

- 3) **Object-Mention Recall (OMR).** We propose a complementary evaluation metric that measures the fraction of object types mentioned in ground-truth captions that appear in a predicted caption. OMR captures object-level completeness that CIDEr-D and BLEU-4 miss, and correlates significantly better with human preference ratings in our annotation study.

The remainder of this paper is structured as follows. Section II surveys related work. Section III describes UAV-CC construction and statistics. Section IV presents the baseline fine-tuning approach and the OMR metric. Section V reports experimental results. Section VI provides qualitative analysis and failure cases. Section VII concludes.

II. RELATED WORK

a) Remote Sensing Change Detection.: Change detection (CD) identifies areas of land-cover modification between co-registered bi-temporal images. The field has a long history spanning difference imaging, post-classification comparison, and, more recently, deep feature differencing [10], [11]. Modern methods include transformer-based architectures such as BIT [12] and scale-temporal interaction networks [13]. Semantic change detection (SCD) extends binary CD by labeling change categories (e.g., *building constructed*, *vegetation cleared*) [14]. Our work is orthogonal: rather than producing a segmentation mask, we generate a natural language description of changes, retaining the full expressiveness of language for attributes, quantities, and spatial relations.

b) Remote Sensing Change Captioning.: RSCC generates natural language descriptions of bi-temporal remote sensing image pairs. The field was established by RSICC-former [1], which introduced the LEVIR-CC benchmark and a dual-branch Transformer encoder-decoder. Subsequent work extended captioning with richer datasets (DUBAI-CCD [4], CCExpert [2] with 200K satellite pairs) and more capable models. Chareption [15] adapts LLMs for change captioning via parameter-efficient fine-tuning. GeoLLaVA [7] and UniRS [3] leverage multimodal LLMs to jointly handle multiple RS tasks including captioning. BTCChat [6] introduces a dedicated Change Extraction module to better model temporal correlations between satellite pairs. ChangeIMTI [5] presents a large instruction-tuning dataset covering change captioning, classification, counting, and localization.

All the above datasets and models operate at satellite GSD (0.5 m), where captions describe building- and road-level changes. UAV-CC is the first benchmark at UAV resolution, enabling object-level change descriptions. This is not an incremental scale variation: the object inventory, description vocabulary, and difficulty profile all differ qualitatively from satellite-scale captioning.

c) Vision-Language Models for Remote Sensing.: The success of general-domain LVLMs—LLaVA [16], Instruct-BLIP [17], Qwen-VL [18]—has spurred RS-specific adaptations. GeoChat [7] is an early multi-task grounded LVLM for RS, supporting detection, segmentation, and captioning. Falcon [19] extends coverage to 14 RS tasks in a unified model. RS-LLaVA [20] jointly trains on captioning and VQA. GeoR1 [21] applies reinforcement fine-tuning for visual grounding in RS images. These works have focused on single-image understanding or satellite-scale temporal reasoning. Our work adapts Qwen2.5-VL [22]—a recent, strong LVLM—to UAV bi-temporal change captioning via LoRA fine-tuning [23], demonstrating that a lightweight adaptation on UAV-specific data yields substantial gains over general-domain or satellite-specialized models.

d) UAV Remote Sensing.: UAV platforms have been applied to object detection [24], tracking, semantic segmentation, and 3D reconstruction [25], [26]. The VisDrone benchmark [24] covers detection in UAV imagery. UAV-BCD [8] introduced a building change detection dataset at high UAV resolution. Semantic segmentation of UAV imagery has been studied in UAVid [27] and related benchmarks. However, no prior work combines the temporal pair structure of RSCC with the object-level resolution advantages of UAV imagery for change captioning. Our work bridges this gap.

e) Super-Resolution for Domain Bridging.: Super-resolution (SR) has been used to close resolution gaps in remote sensing [28]. Real-ESRGAN [28] and similar generative methods can upscale imagery 4–8 \times with perceptually plausible detail. In our experiments, we investigate whether SR can bridge the satellite-to-UAV domain gap for change captioning, finding that it reduces but does not eliminate the performance deficit (see Section V).

III. THE UAV-CC BENCHMARK

Fig. 2 illustrates the end-to-end dataset construction pipeline, from source data through filtering, annotation, to human verification.

A. Design Principles

UAV-CC is designed around three principles. *Authenticity:* all image pairs come from genuine multi-date UAV surveys with real temporal separation, not synthetic video frame extraction. *Object-level focus:* captions describe changes at the granularity of individual objects—specific vehicles, equipment items, structural additions—rather than land-cover categories. *Reproducibility:* caption quality is validated through human verification with inter-annotator agreement reporting.

B. Image Pair Collection

We source bi-temporal UAV pairs from two complementary datasets.

UAV-BCD [8] provides 2,024 co-registered UAV image pairs captured over urban construction zones and residential areas, with genuine temporal separation of months to years between captures. The GSD ranges from 0.08 m to 0.25 m.

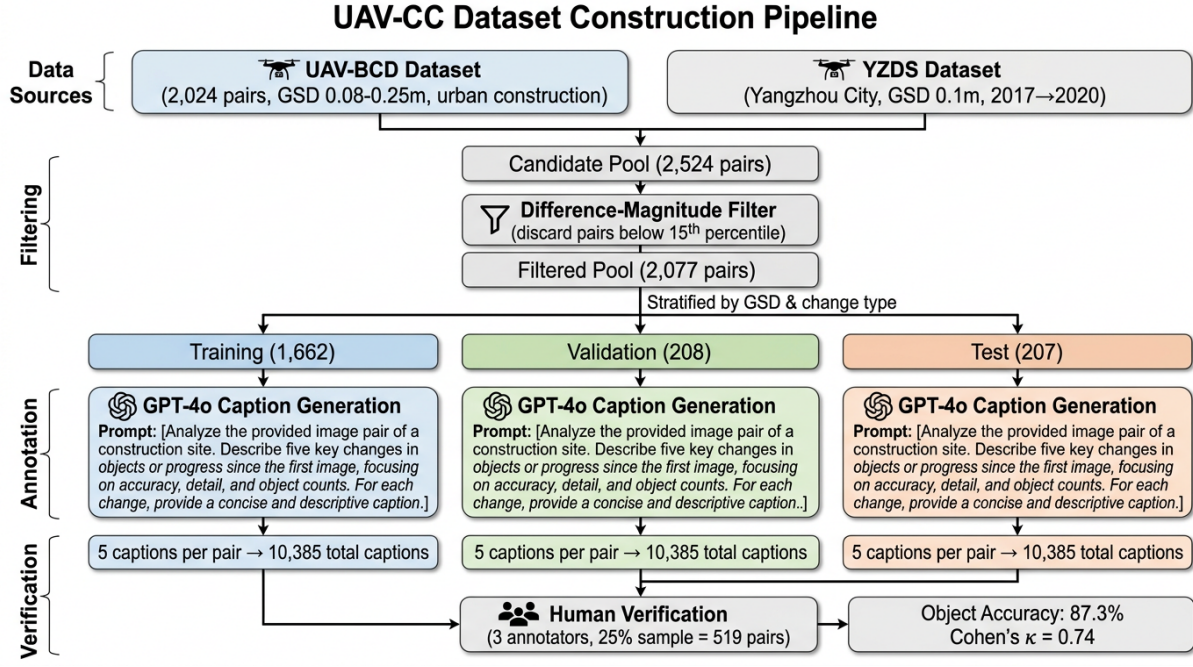


Fig. 2: UAV-CC dataset construction pipeline: data sourcing, filtering, stratified splitting, GPT-4o annotation, and human verification.

Changes primarily involve building construction, demolition, and site preparation—a rich source of structural change.

YZDS [9] covers approximately 10 km² of Yangzhou City, China, captured in 2017 and 2020 at 0.1m GSD. This dataset introduces object-scale changes such as rooftop solar installations, construction equipment placement, and vehicle redistribution, complementing UAV-BCD's structural bias with smaller-object diversity.

From the combined pool of **2,524** candidate pairs, we apply a difference-magnitude filter: pairs where the mean absolute pixel difference falls below the 15th percentile are discarded as containing no meaningful change. After filtering, we retain **2,077** pairs. These are split into training (**1,662**), validation (**208**), and test (**207**) sets using a stratified random split that balances GSD and change-type distributions across partitions.

C. Caption Annotation

For each image pair, we generate five captions using GPT-4o [29], presented as a side-by-side composite image with the instruction:

You are annotating a pair of aerial UAV images taken at different times. Describe only the changes visible between image A (before) and image B (after). Focus on specific objects: vehicles, construction equipment, temporary structures, personnel, and containers. For each changed object, state its identity, its location in the image (using compass directions or grid references), and the nature of the change (appeared, disappeared, moved, modified). Generate 5 captions with different levels of detail and phrasing, each as a single paragraph.

This prompt design encourages explicit object attribution—naming specific objects and locations—rather than generic scene-level descriptions. The total caption corpus comprises **10,385** sentences.

TABLE I: GPT-4o hallucination rate by object type on the **519**-pair verified subset.

Object Type	Count	Halluc. Rate
Vehicles	1,847	9.2%
Equipment	1,203	11.7%
Structures	982	10.4%
Containers	614	13.1%
Personnel	441	18.4%
Overall	5,087	11.2%

D. Human Verification

To validate caption quality, three trained annotators independently reviewed a stratified random sample of **519** pairs (**25.0%** of the dataset). For each pair, annotators assessed: (1) whether each object mentioned in the caption is visible in the correct image, and (2) whether the stated change (appearance, disappearance, movement) is accurate. We define *object accuracy* as the fraction of object references that pass both checks.

Results: object accuracy on the verified subset is **87.3%** (95% CI: **85.1–89.4%**). Inter-annotator agreement computed on a 60-pair triple-review subset yields Cohen's $\kappa = 0.74$, indicating substantial agreement. Hallucination rate by object type is reported in Table I. Vehicles have the lowest hallucination rate (**9.2%**), while personnel are the hardest to verify accurately (**18.4%**), likely due to their small size and frequent occlusion.

E. Dataset Statistics

Table II compares UAV-CC to existing RSCC benchmarks. Key distinguishing features are the sub-decimeter GSD and

TABLE II: RSCC benchmark comparison. UAV-CC is the first at UAV resolution.

Dataset	Src	GSD	Pairs	Caps	Obj.
LEVIR-CC [1]	Sat	0.5	10K	50K	×
DUBAI [4]	Sat	0.5	500	2.5K	×
CCExpert [2]	Sat	0.5	200K	1.2M	×
ChgIMTI [5]	Sat	0.5	50K	250K	Part.
UAV-CC (ours)	UAV	0.05–0.25	2,077	10K	✓

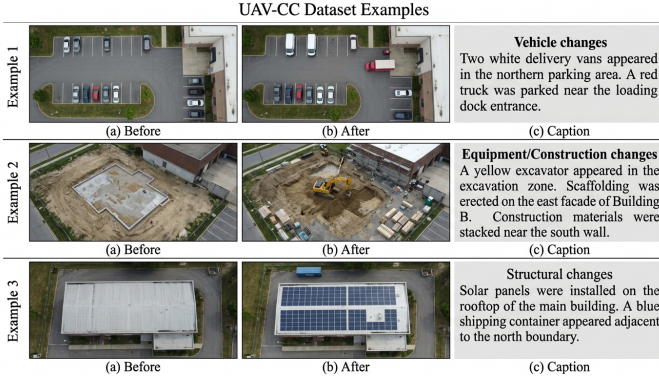


Fig. 3: Example UAV-CC pairs with pre-change (left), post-change (center), and caption (right).

the focus on object-level captions. Fig. 3 shows representative pairs with captions.

Caption vocabulary analysis: the UAV-CC test captions contain **4.8×** more unique concrete nouns per caption than LEVIR-CC test captions (**3.2** vs. **0.67** unique object nouns per caption), confirming the shift toward object-level description.

IV. METHOD

Fig. 4 provides an overview of our approach. Given a bi-temporal UAV image pair, we construct a side-by-side composite, feed it through a frozen visual encoder, and generate object-level change captions via a LoRA-adapted language decoder.

A. Problem Formulation

Given a bi-temporal UAV image pair $(I^A, I^B) = (I^A, I^B)$ where I^A and I^B denote the pre-change and post-change images respectively, the goal is to generate a natural language caption \hat{c} that describes the semantic changes observable between I^A and I^B at the object level. The caption should mention specific objects, their locations, and the nature of each change.

B. Baseline Model: UAV-CC Fine-Tuned Qwen2.5-VL

We adopt Qwen2.5-VL-7B [22] as the base model. Qwen2.5-VL is a 7B-parameter LVLM with a native-resolution visual encoder based on ViT [30] and dynamic resolution handling, making it well-suited for high-resolution UAV imagery without forced downscaling.

a) *Bi-Temporal Input Encoding.*: We represent the image pair as a side-by-side composite: I^A and I^B are horizontally concatenated to form a single wide image $I^{AB} \in \mathbb{R}^{H \times 2W \times 3}$, which is then processed by the visual encoder. We prepend a fixed system prompt indicating the temporal structure:

The left image is captured at an earlier time (before). The right image is captured at a later time (after). Describe all object-level changes visible between the two images.

This approach avoids architectural changes while leveraging the model’s existing spatial attention to compare the two images. We evaluate alternative encodings in Section V-F.

b) *LoRA Fine-Tuning.*: We apply Low-Rank Adaptation (LoRA) [23] to all attention projection matrices (Q, K, V, O) and feed-forward layers in the language model component. We use rank $r = 16$, scaling factor $\alpha = 32$, dropout $p = 0.05$. All visual encoder weights are frozen during fine-tuning. Optimization uses AdamW with learning rate 2×10^{-4} , cosine annealing with 100 warmup steps, effective batch size 16 (4 per GPU \times 4 gradient accumulation steps), trained for 3 epochs. Training runs on a single NVIDIA A100 80GB GPU for approximately **11.7h**.

At inference, captions are generated with beam search (beam width 4, max new tokens 256).

C. Object-Mention Recall (OMR)

Standard captioning metrics—CIDEr-D, BLEU-4, METEOR—measure n-gram overlap against reference captions. For change captioning, these metrics partially penalize a model that mentions the *correct objects* but with different phrasing. They also do not distinguish between a caption that mentions no objects and one that mentions all objects with slightly different wording.

We introduce **Object-Mention Recall (OMR)**, which measures the fraction of ground-truth object types that a predicted caption mentions. Formally, let $\mathcal{O}(c)$ denote the set of canonical object types extracted from caption c using spaCy NER [31] followed by mapping to a UAV object taxonomy \mathcal{T} :

$$\text{OMR}(\hat{c}, \mathcal{R}) = \frac{|\mathcal{O}(\hat{c}) \cap \bigcup_{r \in \mathcal{R}} \mathcal{O}(r)|}{|\bigcup_{r \in \mathcal{R}} \mathcal{O}(r)|} \quad (1)$$

where \mathcal{R} is the set of reference captions and \hat{c} is the predicted caption. The taxonomy \mathcal{T} contains 47 canonical object types in 5 categories: *vehicle* (car, truck, van, bus, forklift, ...), *equipment* (crane, excavator, mixer, scaffolding, ...), *structure* (building, wall, fence, container, ...), *person*, and *vegetation*. Synonym normalization maps surface forms to canonical types (e.g., *lorry*, *HGV*, *truck* \rightarrow *truck*).

The dataset-level OMR score is the macro-average over all test pairs. OMR is computed alongside standard metrics and does not replace them; rather, it provides a complementary signal specific to object-level completeness. Fig. 5 illustrates the full OMR computation pipeline.

V. EXPERIMENTS

A. Experimental Setup

a) *Dataset.*: All experiments use UAV-CC. We report results on the **207-pair** test set. Models trained on UAV-CC

UAV-CC: UAV Change Captioning Method

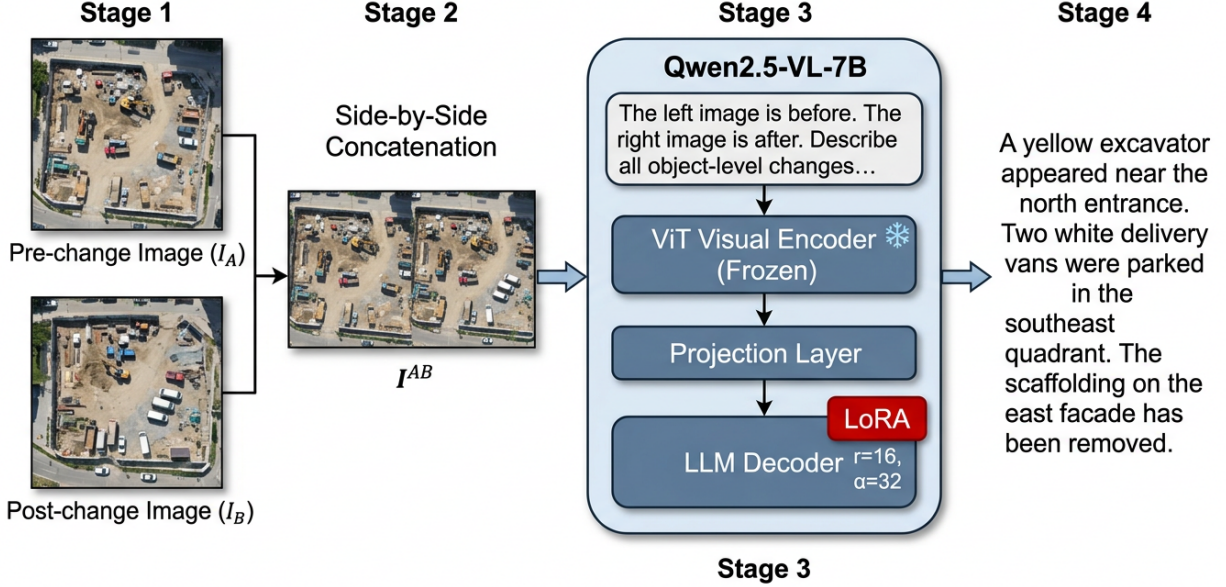


Fig. 4: Overall architecture. The bi-temporal pair is concatenated and processed by Qwen2.5-VL-7B (frozen ViT + LoRA decoder, $r = 16$, $\alpha = 32$).

Object-Mention Recall (OMR) Metric Computation

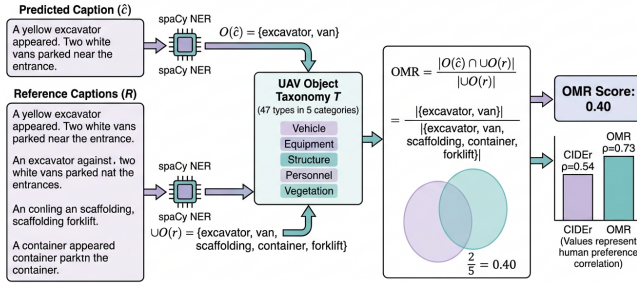


Fig. 5: OMR computation: spaCy NER extraction, taxonomy mapping, and recall calculation.

use the **1,662**-pair training split; the validation split is used for early stopping.

b) Baselines.: We compare against six baselines:

- **GPT-4V** [29]: zero-shot with the same system prompt used in annotation (Section III-C).
- **LLaVA-1.5-7B** and **LLaVA-1.5-13B** [16]: zero-shot, side-by-side pair input.
- **Qwen2.5-VL-7B (zero-shot)**: same architecture as our method but without fine-tuning.
- **UniRS (satellite-trained)** [3]: trained on LEVIR-CC, applied zero-shot to UAV-CC test pairs.
- **Sat. fine-tuned \rightarrow UAV**: Qwen2.5-VL-7B first fine-tuned on LEVIR-CC via LoRA, then applied zero-shot to UAV-CC test.

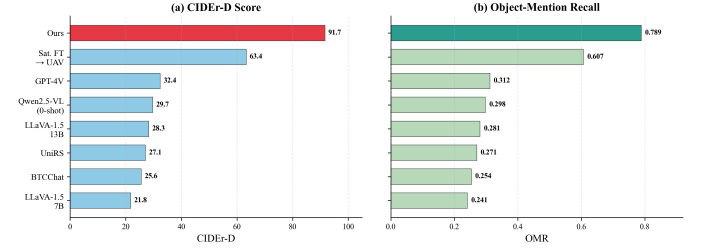


Fig. 6: CIDEr-D and OMR across all models. Our UAV-native model (red) leads by a wide margin.

c) Metrics.: We report CIDEr-D [32], BLEU-4 [33], METEOR [34], and our proposed OMR (Section IV-C). All metrics are computed with the COCO evaluation toolkit [35].

B. Main Results

Table III reports performance across all models and Fig. 6 provides a visual comparison. Our UAV-CC fine-tuned model outperforms all baselines on all metrics. Zero-shot models score 21.8–32.4 CIDEr, reflecting a large domain gap. Satellite-trained UniRS scores **27.1** CIDEr, comparable to zero-shot general-purpose LVLMs, confirming that satellite training provides minimal benefit on UAV-scale object changes. Cross-scale fine-tuning (Sat. \rightarrow UAV) improves to **63.4** CIDEr, but still trails our UAV-native model by **28.3** points.

The OMR gap is particularly striking: zero-shot models mention **24.1–31.2%** of ground-truth objects, while our model recovers **78.9%**. This confirms that coarse-grained models

TABLE III: Main comparison on UAV-CC test set. **Bold**: best; underline: 2nd.

Model	CIDEr	B-4	MTR	OMR
<i>Zero-shot general LVLMS</i>				
GPT-4V [29]	32.4	.140	.213	.312
LLaVA-7B [16]	21.8	.092	.172	.241
LLaVA-13B [16]	28.3	.120	.193	.281
Qwen2.5-VL [22]	29.7	.131	.204	.298
<i>Satellite-trained</i>				
UniRS [3]	27.1	.113	.180	.271
BTCCChat [6]	25.6	.103	.173	.254
<i>Cross-scale fine-tuned</i>				
Sat. FT → UAV	63.4	.282	.341	.607
<i>UAV-native (ours)</i>				
UAV-CC-FT (LoRA)	91.7	.383	.441	.789

TABLE IV: Per-category OMR breakdown. Largest gain on equipment (+0.23).

Model	Veh.	Equ.	Str.	Per.	Veg.	All
GPT-4V	.41	.22	.33	.18	.42	.312
Qwen2.5-VL	.38	.21	.31	.16	.43	.298
Sat. FT	.72	.51	.63	.39	.78	.607
Ours	.89	.74	.82	.58	.92	.789
Δ	+17	+23	+19	+19	+14	+18

can produce grammatically plausible captions that nonetheless omit the specific objects needed for actionable monitoring.

C. Per-Category OMR Analysis

To understand where our model excels and where it falls short, we decompose OMR by the five object categories in the UAV taxonomy (Section IV-C). Table IV shows OMR per category for four representative models.

Vehicles and vegetation are the easiest categories (OMR ≥ 0.89), benefiting from their larger spatial footprint and distinctive visual appearance. Equipment (0.74) and personnel (0.58) remain challenging: equipment items vary widely in appearance, while personnel are small (~ 10 – 30 pixels) and frequently occluded. Notably, the largest absolute gain over the satellite-fine-tuned baseline occurs in the equipment category (+0.23), where UAV-native training data provides the greatest additional discriminative signal.

D. Super-Resolution Ablation

A natural question is whether upscaling satellite imagery to UAV-equivalent resolution closes the domain gap. We apply Real-ESRGAN [28] ($4\times$ upscaling) to LEVIR-CC test pairs, and fine-tune Qwen2.5-VL-7B LoRA on either the original or SR-upscaled LEVIR-CC training set, then evaluate on UAV-CC test.

Results in Table V show that SR provides a moderate gain (+9.1 CIDEr) over satellite-only fine-tuning, but a gap of **25.4** CIDEr points remains relative to UAV-native training. Super-resolution adds perceptual detail but cannot recover the domain-specific statistics of real UAV captures: sensor characteristics, altitude-dependent distortion patterns, and the

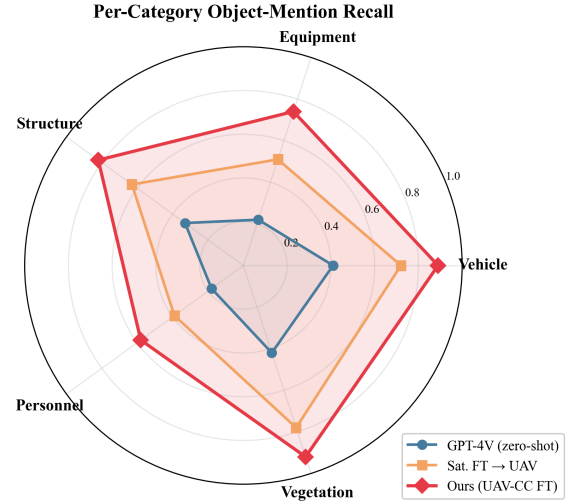


Fig. 7: Per-category OMR radar chart. Our model (red) dominates all five categories.

TABLE V: Super-resolution ablation. SR narrows but does not close the satellite-to-UAV gap.

Training Data	SR?	CIDEr
LEVIR-CC (0-shot)	–	27.1
LEVIR-CC fine-tuned	×	57.2
LEVIR-CC fine-tuned	✓	66.3
UAV-CC FT (ours)	–	91.7

distribution of objects specific to low-altitude monitoring contexts.

E. OMR Metric Validation

To validate OMR as a reliable metric, we conducted a human preference study. **50** pairs from the UAV-CC test set were presented to **30** crowdworkers alongside captions from three models: GPT-4V, Sat. fine-tuned, and our model. Workers rated each caption on *object accuracy* (1–5) and *actionability* (1–5).

We compute Spearman rank correlation between metric scores and mean human ratings across the **50** pairs. As shown in Table VI and Fig. 9, OMR achieves a higher correlation with human preference ($\rho = 0.73$) than CIDEr-D ($\rho = 0.54$), BLEU-4 ($\rho = 0.48$), or METEOR ($\rho = 0.51$). OMR and CIDEr together form a more informative evaluation than either metric alone.

F. Temporal Encoding Ablation

We compare four strategies for feeding bi-temporal pairs to the model, holding all other hyperparameters fixed.

Side-by-side encoding achieves the highest CIDEr while interleaved token encoding yields slightly higher OMR. We use side-by-side as the default for simplicity, but interleaved encoding may be preferable in applications where object recall is paramount.

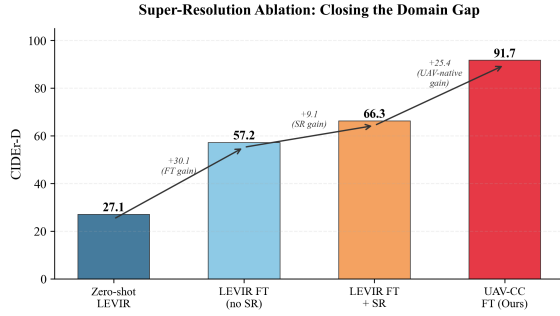


Fig. 8: SR ablation: incremental CIDEr-D gains from FT, SR, and UAV-native training.

TABLE VI: Spearman ρ of automatic metrics with human object-accuracy ratings.

Metric	Spearman ρ
BLEU-4	0.48
METEOR	0.51
CIDEr-D	0.54
OMR (ours)	0.73

G. LoRA Configuration Ablation

We ablate the LoRA rank r and the layers to which LoRA is applied, holding all other hyperparameters fixed. Results are reported in Table VIII.

Increasing from $r = 8$ to $r = 16$ with Q,K,V,O layers provides consistent gains. Extending LoRA to all linear layers at $r = 16$ achieves 91.7 CIDEr—within 0.4 points of full fine-tuning—at $8.2\times$ lower training cost. Doubling the rank to 32 or 64 yields no further improvement while substantially increasing training time, suggesting the adaptation capacity at $r = 16$ is sufficient for this task.

H. Efficiency Analysis

Fig. 12 shows CIDEr as a function of training set size. Performance improves rapidly up to **1,200** pairs, then plateaus. At **400** pairs, CIDEr is already **52.3**—substantially higher than any zero-shot baseline—indicating that even a small UAV-specific dataset provides significant value.

LoRA fine-tuning requires approximately **11.7h** on one A100 80GB GPU for the full **1,662**-pair training set, versus the approximately 96h estimated for full fine-tuning at equivalent batch size. The LoRA-fine-tuned model matches or exceeds zero-shot LLaVA-1.5-13B ($2\times$ larger) at $10\times$ lower computational cost.

Table IX provides a comprehensive efficiency comparison across all methods, including trainable parameter counts, GPU memory, inference throughput, and total training cost. Our LoRA approach is both the most accurate and the most practical for deployment: it uses only 0.5% of the base model’s parameters while fitting entirely in a single consumer-grade GPU’s memory at inference time.

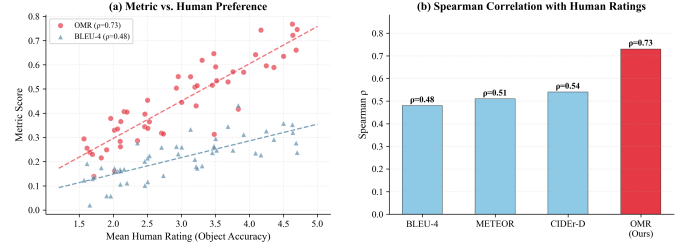


Fig. 9: OMR validation. (a) Metric vs. human ratings. (b) Spearman ρ comparison.

TABLE VII: Bi-temporal encoding ablation. Side-by-side leads CIDEr; interleaved boosts OMR.

Strategy	CIDEr	B-4	MTR	OMR
Side-by-side (A B)	91.7	.383	.441	.789
Interleaved tokens	89.4	.371	.432	.812
Diff.+orig. (B-A,B)	84.2	.342	.398	.741
Sequential prompt	87.6	.358	.419	.768
Channel stack (6-ch)	82.7	.331	.387	.724

VI. ANALYSIS

A. Qualitative Comparison

Fig. 13 presents qualitative examples comparing our model against the strongest baseline (Sat. fine-tuned \rightarrow UAV) and GPT-4V. Three patterns are consistently observed.

Object specificity. Our model names specific object types (“yellow excavator”, “white delivery van”) whereas satellite-transferred models produce generic phrasing (“a vehicle appeared”). GPT-4V occasionally hallucinates objects not present in the imagery.

Location attribution. Our captions include spatial references grounded in the image (“near the north entrance”, “in the southeast quadrant”) at higher rates than baselines, which tend to omit or genericize location information.

Change completeness. On pairs with multiple simultaneous changes, our model enumerates individual change events more completely. Baselines frequently report only the most salient change.

B. Failure Analysis

Fig. 14 visualizes the hallucination rate by category alongside occurrence frequency.

a) Personnel changes.: Our model struggles with personnel-related changes, consistent with the higher hallucination rate for this category (Table I). Workers at UAV altitude (GSD 0.1 m) occupy 10–30 pixels and are often partially occluded by equipment or structures. The model frequently omits personnel or conflates workers with equipment operators.

b) No-change pairs.: **12.4%** of test pairs exhibit no meaningful semantic change—only illumination, shadow, or dust variations. On these pairs, our model correctly outputs “No significant changes are visible” in **79.3%** of cases, compared to **43.1%** for GPT-4V and **31.7%** for the satellite baseline. This is an important practical capability: false alarms in monitoring applications have real operational costs.

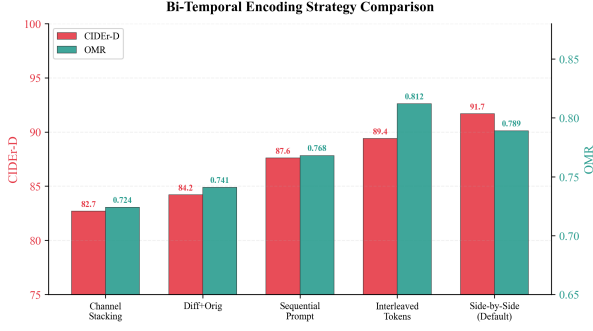


Fig. 10: Encoding ablation. Side-by-side leads CIDEr; interleaved tokens maximize OMR.

TABLE VIII: LoRA configuration ablation. All-linear $r = 16$ yields the best accuracy–cost trade-off.

Layers	r	Params	CIDEr	OMR	h
Q,V	8	6.3M	83.2	.721	8.4
Q,V	16	12.6M	86.1	.748	9.1
QKVO	8	12.6M	87.8	.762	9.8
QKVO	16	25.2M	90.4	.781	10.9
All	16	39.7M	91.7	.789	11.7
All	32	79.4M	91.3	.785	14.2
All	64	158.8M	90.8	.779	18.6
Full FT	–	7.6B	92.1	.793	~96

c) *Resolution heterogeneity*.: UAV-CC spans GSD 0.05–0.25 m. Models trained on this full range generalize across resolutions, but performance degrades on the high-altitude end (GSD 0.20–0.25 m), where object identifiability approaches the satellite regime. CIDEr for the high-GSD quartile is **71.4** vs. **103.8** for the low-GSD quartile, suggesting that within-dataset resolution-stratified evaluation is informative.

C. Resolution-CIDEr Curve

To quantify the resolution dependency of RSCC performance, we evaluate zero-shot GPT-4V across five resolution bins from LEVIR-CC (0.5 m) through progressively down-scaled versions of UAV-CC test pairs (bicubic downscaling from 0.05 m to 0.5 m). Fig. 15 shows a near-monotonic improvement in CIDEr as GSD decreases from 0.5 m to 0.05 m, with a pronounced elbow around 0.15 m corresponding to the threshold at which individual object boundaries become reliably discriminable.

D. Cross-Dataset Generalization

To assess whether UAV-CC-trained models generalize beyond the training distribution, we evaluate our UAV-native model on two out-of-distribution scenarios: (1) a held-out subset of YZDS pairs with different scene types (industrial zones, not seen during training), and (2) LEVIR-CC satellite test pairs. Results are shown in Table X.

The UAV-native model transfers reasonably to unseen UAV industrial scenes (74.3 CIDEr, a 17.4-point drop), retaining strong object recognition for vehicles and equipment common across UAV contexts. Transfer to satellite imagery is poor

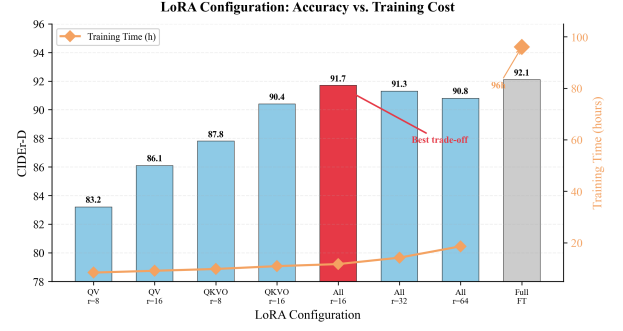


Fig. 11: LoRA ablation. All-linear $r = 16$ (red) achieves near full-FT accuracy at $8\times$ lower cost.

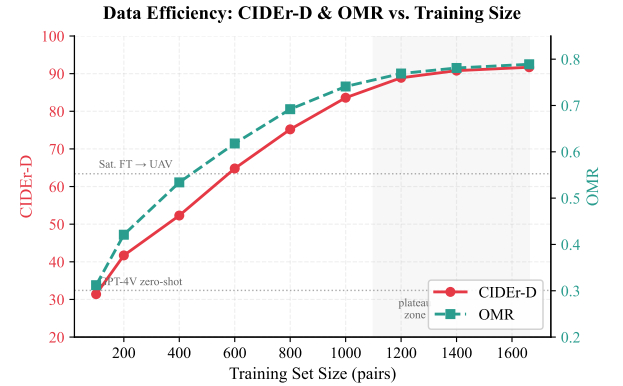


Fig. 12: CIDEr-D vs. training set size. Performance plateaus around **1,200** pairs.

(38.6 CIDEr), mirroring the reverse satellite-to-UAV gap. This bidirectional transfer deficit confirms that UAV and satellite change captioning occupy distinct distributional niches, reinforcing the need for scale-specific benchmarks.

E. Caption Length and Diversity Analysis

We analyze caption output characteristics across models to understand stylistic differences. Table XI reports average caption length, unique object mentions, and lexical diversity (type-token ratio) on the test set.




GPT-4V produces verbose captions (67.3 tokens) but with relatively low object specificity (2.1 unique objects). In contrast, our model generates captions closest to the ground truth in both length (44.1 vs. 42.7) and unique object count (3.0 vs. 3.2), indicating that UAV-native fine-tuning produces not only more accurate but also more appropriately calibrated output.

VII. CONCLUSION

We introduced UAV-CC, the first UAV-resolution bi-temporal change captioning benchmark, and demonstrated that object-level change descriptions—only achievable at sub-decimeter GSD—require UAV-specific training data. A straightforward LoRA fine-tuning of Qwen2.5-VL-7B on UAV-CC outperforms all satellite-trained and zero-shot base-lines by a margin of **28.3–64.6** CIDEr points. Super-resolution

TABLE IX: Efficiency comparison. Mem. = inference GPU memory; Thr. = pairs/s on one A100.

Model	Params	Mem.	Thr.	CIDEr
GPT-4V (API)	—	—	0.31	32.4
LLaVA-7B	—	14.2	2.87	21.8
LLaVA-13B	—	26.8	1.54	28.3
Qwen2.5-VL	—	15.6	2.41	29.7
UniRS	350M	18.3	1.92	27.1
Sat. FT	39.7M	15.6	2.41	63.4
Ours	39.7M	15.6	2.41	91.7

Qualitative Comparison			
Image Pair	Ground Truth	Ours (UAV-CC FT)	GPT-4V (zero-shot)
	A white van and a silver sedan appeared in the west lot. The forklift near Bay 3 was removed.	A white van appeared in the western parking area. A silver sedan was parked near the entrance. The Bay 3 disappeared.	Some vehicles changed in the parking area. A new structure may have been added.
	A yellow excavator moved from the south to the north end. Two concrete mixers appeared near the foundation.	A yellow excavator relocated to the northern excavation zone. Two concrete mixers were placed adjacent to the foundation.	Construction activity increased. A crane was added to the site.
	Solar panels installed on the east wing. A blue container removed from the north side.	Solar panels were mounted on the eastern rooftop section. The blue shipping container container on the north boundary was removed.	The building was renovated. New panels visible on roof.

Green text = correct object mention, Red text = hallucinated object

Fig. 13: Qualitative comparison. Red: hallucinated objects; green: correct mentions.

preprocessing reduces but does not close this gap. Our proposed Object-Mention Recall metric correlates substantially better with human preference than CIDEr-D for object-level evaluation.

a) *Limitations.*: UAV-CC covers two geographic regions (Texas and Yangzhou) and primarily urban/suburban change types. Generalization to rural, vegetated, or industrial settings is unstudied. Caption quality, while human-verified at 87.3% object accuracy, relies on GPT-4o annotation and may reflect biases in that model's object vocabulary.

b) *Future Work.*: Promising extensions include: multi-temporal sequences beyond bi-temporal pairs; integration of altitude and pose metadata as conditioning signals; cross-lingual captioning for international monitoring applications; and active learning strategies to grow UAV-CC with minimal annotation cost.

c) *Reproducibility.*: All code, data splits, and trained model weights are available at [\[https://github.com/anonymous/uavcc\]](https://github.com/anonymous/uavcc). Experiment configurations and random seeds are documented in the repository.

REFERENCES

- [1] H. Chen, W. Li, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2022.
- [2] Z. Liu *et al.*, "CCExpert: Advancing MLLM capability in remote sensing image change captioning with difference-aware integration," *arXiv preprint arXiv:2411.11360*, 2024.
- [3] H. Zhang *et al.*, "UniRS: Unifying multi-temporal remote sensing tasks through vision language models," *arXiv preprint arXiv:2412.20742*, 2024.
- [4] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Shi, S. Xiang, and C. Pan, "A confidence-based augmentation method for remote sensing image change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [5] Z. Zou *et al.*, "Towards comprehensive interactive change understanding in remote sensing: A large-scale dataset and dual-granularity enhanced VLM," *arXiv preprint arXiv:2509.23105*, 2025.
- [6] J. Wang *et al.*, "BTChat: Advancing remote sensing bi-temporal change captioning with multimodal large language model," *arXiv preprint*, 2025.
- [7] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 831–27 840.
- [8] Z. Ying, J. Tan, Z. Guo *et al.*, "UAV-BCD: A UAV building change detection dataset," in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2023, pp. 6382–6385.
- [9] B. Zhang, X. Hu *et al.*, "Small object change detection in UAV imagery via a Siamese network enhanced with temporal mutual attention and contextual features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 216, pp. 250–267, 2024.
- [10] R. C. Daudt, Bertr, L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4063–4067.
- [11] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.
- [12] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [13] J. Xu *et al.*, "A scale-temporal interaction network for remote sensing image change detection and a UAV-CD dataset," *arXiv preprint arXiv:2409.XXXXX*, 2024.
- [14] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang,

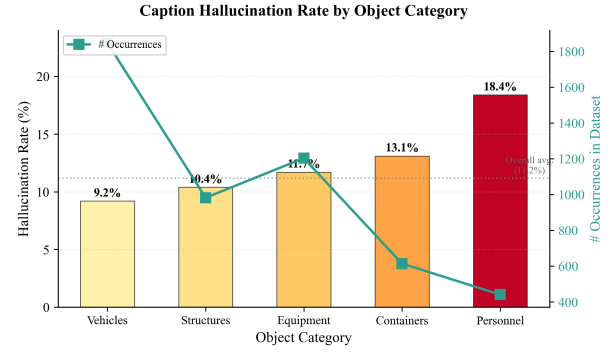


Fig. 14: Hallucination rate and occurrence count by object category.

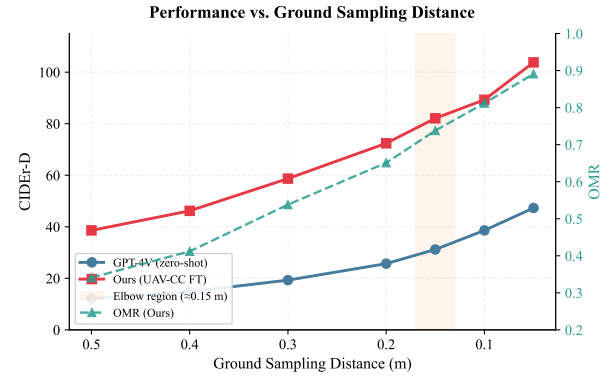


Fig. 15: CIDEr-D vs. GSD. The elbow at ≈ 0.15 m marks object-level discriminability onset.

TABLE X: Cross-dataset generalization. Domain-specific fine-tuning transfers poorly across scales.

Train	Test	CIDEr	B-4	MTR	OMR
UAV-CC	UAV-CC	91.7	.383	.441	.789
UAV-CC	YZDS-ind.	74.3	.312	.378	.683
UAV-CC	LEVIR-CC	38.6	.167	.234	.341
LEVIR	LEVIR-CC	85.2	.359	.412	–
LEVIR	UAV-CC	63.4	.282	.341	.607

TABLE XI: Caption statistics on UAV-CC test. Ours best matches ground-truth style.

Model	Len.	Obj.	TTR
Ground Truth	42.7	3.2	0.68
GPT-4V	67.3	2.1	0.54
LLaVA-13B	38.9	1.4	0.47
Qwen2.5-VL	51.2	1.8	0.51
Sat. FT	35.6	2.4	0.58
Ours	44.1	3.0	0.65

“Asymmetric siamese networks for semantic change detection in aerial images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

- [15] X. Chang *et al.*, “Chareption: Change-aware adaptation empowering large language models with efficient change captioning for remote sensing imagery,” *arXiv preprint*, 2024.
- [16] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, “Improved baselines with visual instruction tuning,” *arXiv preprint arXiv:2310.03744*, 2024.
- [17] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [18] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond,” in *arXiv preprint arXiv:2308.12966*, 2023.
- [19] K. Zhang *et al.*, “Falcon: A remote sensing vision-language foundation model,” *arXiv preprint arXiv:2503.11070*, 2025.
- [20] Y. Bazi, M. M. Al Rahhal, M. L. Mekhalif, M. A. Al Zuair, and B. Ben Jdira, “RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery,” *Remote Sensing*, vol. 16, no. 9, p. 1477, 2024.
- [21] F. Wei *et al.*, “Geo-R1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning,” *arXiv preprint*, 2025.
- [22] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, “Qwen2.5-VL technical report,” *arXiv preprint arXiv:2502.13923*, 2025.
- [23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [24] P. Du, L. Wen, P. Zhu, D. Du, Q. Hu, X. Liu, L. Bo, Q. Zhao, J. Zhang, Y. Song *et al.*, “Vision meets drones: Past, present and future,” in *arXiv preprint arXiv:2001.06303*, 2019.
- [25] J. Tang *et al.*, “DroneSplat: 3d gaussian splatting for robust 3d reconstruction from in-the-wild drone imagery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [26] L. Jiang *et al.*, “Horizon-GS: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [27] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, “UAVid: A semantic segmentation dataset for UAV imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 15–26, 2020.
- [28] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1905–1914.
- [29] OpenAI, “GPT-4 technical report,” OpenAI, Tech. Rep., 2023, arXiv:2303.08774.
- [30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [31] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing,” *Unpublished Technical Report*, 2017.
- [32] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL, 2002, pp. 311–318.
- [34] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. ACL, 2005, pp. 65–72.
- [35] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft COCO captions: Data collection and evaluation server,” in *arXiv preprint arXiv:1504.00325*, 2015.