C3-Bench: A Context-Aware Change Captioning Benchmark

Jae-Woo Kim[1], Hyeongbeom Kim[1], and Ue-Hwan Kim[1,2*]

[1]Gwangju Institute of Science and Technology, Gwangju, Republic of Korea
[2]GIST InnoCORE AI-Nano Convergence Institute for Early Detection of Neurodegenerative Diseases, Gwangju Institute of Science and Technology, 61005 Gwangju, Republic of Korea
kjw01124@gm.gist.ac.kr, hbk08101@gm.gist.ac.kr, uehwan@gist.ac.kr

Project Page: https://github.com/AutoCompSysLab/C3-Bench

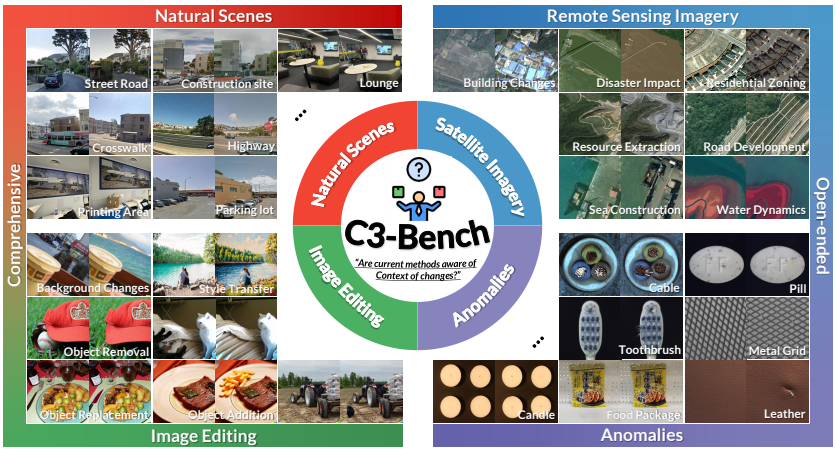

Fig. 1: Overview of C3-Bench. The examples are from each context in C3-Bench.

Abstract. While Change Captioning systems have garnered substantial attention to respond to our evolving world, their true performance on *diverse real-world change contexts* remains largely unexplored due to the lack of comprehensive evaluation frameworks. To fill this gap, we propose **C3-Bench**, a comprehensive benchmark for evaluating **C**ontext-aware **C**hange **C**aptioning. C3-Bench features: (1) 4,996 human-labeled image pairs of 51 real-world change contexts across four domains (e.g., natural scenes, remote sensing imagery, image editing, and anomalies), each with diverse, carefully curated scenarios derived from multiple change-centric communities; and (2) the first LLM-as-Judge evaluation framework in the change captioning task that measure fine-grained dimensions (e.g., correctness, specificity, fluency, and relevance), along with a novel reversibility metric exploring whether models understand changes with symmetric consistency. Based on C3-Bench, we benchmark 32 models—including

* Corresponding author

conventional change captioning models, proprietary Large Multimodal Models (LMMs), and 2B-90B open-source LMMs. We reveal a fundamental blind spot in the prevailing change captioning paradigm: Once the change context departs from training-style regimes, conventional models collapse, and even state-of-the-art LMMs such as GPT-5.2 exhibit systematic domain- and position-dependent errors that distort reliable change understanding. By making these hidden failure modes explicit and measurable, we delineate the next frontier for building generalizable and trustworthy change captioning systems. All codes and datasets are publicly available on the project page.

Keywords: Change Captioning · Dataset & Benchmark

1 Introduction

The world is non-stationary by default—structures emerge and decay [94, 115], objects appear and vanish [47], and environments drift under natural- and human-induced activities [35, 68]. Therefore, perceiving *what has changed* and communicating *how it changed* for intelligent systems to operate outside the controlled laboratories is no longer optional but *indispensable* prerequisite for effective collaboration with humans; advancement here has an immediate impact on various downstream practices [19–21, 24, 70, 74, 106]—including incident monitoring, autonomy, anomaly reporting, remote sensing, and media forensics.

Fig. 2: What has changed? (Motivation) Without a *context*, this generic question can admit multiple logically valid descriptions.

Meanwhile, articulating change is fundamentally challenging due to the inherent ambiguity of "*change*". Consider a single image pair shown in Fig. 2. When one is asked to describe the change between the given image pair, what might first come to mind is "*in which context?*", as the definition of correct change can vary depending on the given ***context***: the valid description would be "*the snow has covered the ground, and cloud cover has decreased.*" in respect of weather, whereas it is "*a train has appeared on the left side of the tracks.*" for railway surveillance, with weather differences treated as pseudo-changes [2, 48]. To meaningfully communicate and determine the correct change description among multiple logically valid alternatives in a heterogeneous visual world, each change must be grounded in *specific contexts* and *associated criteria* which clearly define the underlying semantics.

However, most prevailing benchmarks [45, 47, 52, 60, 75] largely overlook this point and describe changes either arbitrarily or implicitly—hindering a comprehensive understanding of diverse change contexts. Furthermore, their contextual coverage remains confined to narrow datasets—obscuring models' genuine perfor-

Table 1: Comparison with prior evaluation benchmarks. Our C3-Bench covers a broad range of change-centric domains and contexts—providing a robust testbed for change captioning systems. See Fig. 1 for detailed domains included in **ALL**.

Benchmark	Dataset			Metrics		Baselines and Protocols	
	Domain	# Contexts	# Pairs	Open-ended	Reversibility	LMM-support	Open-domain
Synthetic Environment							
CLEVR-Change [60]	-	-	7,950	✗	✗	✗	✗
CLEVR-DC [47]	-	-	4,800	✗	✗	✗	✗
Real-world Environment							
LEVIR-CC [52]	Remote	1	1,920	✗	✗	✗	✗
SpotTheDiff [45]	Natural	1	1,270	✗	✗	✗	✗
IER [75]	Editing	3	495	✗	✗	✗	✗
C3-Bench (Ours)	**ALL**	**51**	**4,996**	✓	✓	✓	✓

mance in real-world scenarios. To bridge these gaps, we introduce **C3-Bench**—a comprehensive benchmark for evaluating **C**ontext-aware **C**hange **C**aptioning. Unlike previous works, C3-Bench offers a broader set of real-world change contexts (e.g, natural disaster, resource extraction, construction monitoring) with clearly-defined criteria, derived from multiple change-centric communities—such as change detection [2], image editing [43], and anomaly detection [30]. As shown in Fig. 1 and Tab. 1, C3-Bench consists of 4,996 image pairs spanning 51 real-world change contexts, organized into four overarching visual domains, each with diverse and carefully designed contexts within each domain. To collect data from disparate domains, we develop an efficient annotation pipeline to produce high-quality human-annotated data—reducing data contamination risks.

In addition, many previous benchmarks rely on fixed reference-based scoring metrics such as BLEU [59] and ROUGE [51], which fail to capture the nuanced semantic difference of change descriptions. To overcome the challenges of evaluating open-ended change captioning, we establish LLM-as-Judge [32] metrics that score fine-grained dimensions of model predictions—facilitating a more human-aligned evaluation that reveals genuine progress toward real-world generalization. Moreover, we evaluate reversibility by swapping the input image order and examining whether the model understands changes in a symmetrically consistent manner—a crucial indicator for system reliability [48]. We believe C3-Bench to lay a solid foundation for developing truly applicable change captioning systems.

Based on C3-Bench, we extensively benchmark 32 models—including conventional change captioning models, proprietary LMMs, and 2B-90B open-source LMMs. Our experiments yield several key findings: (1) Conventional models, despite their reported effectiveness, collapse in open-domain contexts—indicating brittle, dataset-tied notions of what count as change; and (2) While LMMs can close this gap under our criteria prompting, they remain limited by perceptual and spatial failures (Top-2 errors) and order-induced artifacts (pairwise position bias)—calling for more robust and reliable change captioning systems. Furthermore, our ablation study uncovers essential prompting principles for LMM-based change captioning, revealing that explicit conditioning on change criteria, sequential tags, and alignment with canonical temporal order are crucial.

To summarize, our contributions are threefold:

1. **Problem Formulation.** We define the task of Context-aware Change Captioning, a principled and delicate framework that consolidates heterogeneous change-centric problems into a single, unified task formulation.
2. **Benchmark Construction.** We design C3-Bench, a comprehensive benchmark comprising 4,996 human-annotated image pairs spanning 51 real-world change contexts, alongside human-aligned evaluation metrics.
3. **Critical Insights.** We conduct extensive experiments across diverse models, offering the first large-scale analysis of real-world performance and distilling critical insights into where and how research progress should be made.

2 Related Works

Most conventional methods have explored change captioning by enhancing model architectures [15,34,40,42,50,62,64,72,75,77,78,81,104,110] and developing sophisticated training strategies [4, 38, 47, 56, 79, 80, 98, 116]. Their primary focus has been largely centered on improving performance within *a single or a limited set of data-specific change contexts*, whereas the question of how to scale this task to *a broader range of real-world change contexts* remains unanswered. Recently, LMMs such as GPT-4o [1] have shown promising results in understanding changes between a group of images [85,100] or serving as pseudo-annotators [12,39] to generate synthetic change captioning data. However, these efforts formulate change understanding as multiple-choice questions [28,113], rather than requiring models to produce open-ended descriptions of visual changes. Moreover, they do not operate under clearly defined change criteria, relying on generic instructions such as "*describe the difference* [116]." In contrast, our work focuses on open-ended change captioning grounded in a clearly defined context and criteria, enabling a more faithful and structured framework for change understanding.

Contemporary evaluation datasets such as CLEVR [46]-based family [47,60] primarily rely on synthetic tabletop scenes built from toy bricks—yielding overly simplified layouts with limited complexity. Real-world datasets [45, 52, 75] suffer from narrow topical and contextual coverage; LEVIR-CC [52], for example, largely comprises residential zoning over terrestrial surfaces—overlooking other crucial contexts such as natural disasters [35, 111], coastal expansion [71], and land extraction [102]. Furthermore, no existing dataset addresses multiple visual domains simultaneously, due to the context-agnostic problem formulation of existing approaches (see Sec. 3).

Evaluating change descriptions requires capturing nuanced semantic differences beyond surface-level lexical overlaps. However, most conventional metrics [3, 9, 36, 37, 51, 59, 59, 63, 69, 73, 84, 103, 109], rely on fixed reference-based comparisons, failing to capture the semantic equivalence across diverse correct descriptions. This limitation becomes more pronounced in the case of LMMs, whose outputs are flexible, compositional, and often paraphrastic [99]. LLM-based evaluations have emerged as powerful alternatives [32, 54, 55, 83, 92, 117], however, their potential to change captioning remains largely underexplored.

Lastly, although the change captioning task can now be performed by LMMs and is no longer exclusive to conventional models [39,74], most evaluation frameworks still neither incorporate LMMs in a structured manner nor evaluate open-domain generalizability, focusing instead solely on in-domain performance within fixed datasets. To fill these gaps, we systematically evaluate and compare conventional change captioning models and LMMs in a unified framework, C3-Bench, revealing how models generalize across diverse real-world change contexts.

3 Problem Formulation

Given an image pair $\mathcal{I}_A$ and $\mathcal{I}_B$, conventional methods typically formulate change captioning as:

$$\hat{y}_{A\to B} = \arg\max_{y\in\mathcal{Y}} p_\theta\big(y \mid \mathcal{I}_A, \mathcal{I}_B\big), \tag{1}$$

where $\hat{y}_{A\to B}$ denotes the generated change description from $\mathcal{I}_A$ to $\mathcal{I}_B$, $\mathcal{Y}$ indicates the space of all textual descriptions, and θ is the model parameters. However, Eq. (1) overlooks the context-dependent nature of changes. Moreover, it fails to capture the behavior of LMMs, which operate under multimodal conditioning [17, 53, 91] and incorporate textual instructions as part of their input.

To account for this, we reformulate the Eq. (1) as:

$$\hat{y}^c_{A\to B} = \arg\max_{y\in\mathcal{Y}} p_\theta\big(y \mid \mathcal{I}_A, \mathcal{I}_B; \mathcal{C}_c\big), \tag{2}$$

where c indexes the context, $\mathcal{C}_c$ denotes the associated criteria, and $\hat{y}^c_{A\to B}$ represents the model prediction that conforms to the semantics defined by $\mathcal{C}_c$. Here, Eq. (1) can be considered as a special case of Eq. (2) where $\mathcal{C}_c$ is unspecified. LMMs can be explicitly guided on $\mathcal{C}_c$ through textual instructions. When such specifications are absent—as in conventional models or generically prompted LMMs—the $\mathcal{C}_c$ becomes implicit (reduces to $\mathcal{C}_\theta$), reflecting the model's internal biases learned from the training distribution. Based on this formulation, we incorporate conventional models and instruction-guided LMMs within a unified change captioning framework, as well as establish a more comprehensive and controllable basis for change understanding.

4 C3-Bench

4.1 Data Curation

Collecting and labeling change captioning data is inherently challenging due to the costly nature of change annotation [115]. It is particularly difficult to gather and categorize images across diverse contexts while ensuring consistency [97]. To this end, we created C3-Bench with an efficient pipeline described below.

Context Conceptualization. We begin by identifying diverse real-world phenomena where understanding change is crucial. To expand beyond conventional data scopes, we conduct an extensive literature review of diverse change-centric

Table 2: Structure of context-specific change criteria. The criteria consist of three parts: (*i*) change to detect, (*ii*) change to ignore, and (*iii*) tone and nuance. The example below corresponds to the context, *highway*, from the Natural Scenes domain.

Segment	Description
Change to Detect	• Movement, appearance, disappearance, or modification of vehicles, traffic signs, cones, or roadside objects. • Structural or spatial rearrangements of lanes, barriers, or road markings. • Functional or environmental state changes (e.g., traffic signs activated/deactivated).
Change to Ignore	• Lighting or exposure variations • Viewpoint or viewpoint-induced differences • Minor photometric or compression artifacts • Weather, seasonal, or shadow differences
Tone and Nuance	Neutral and objective.

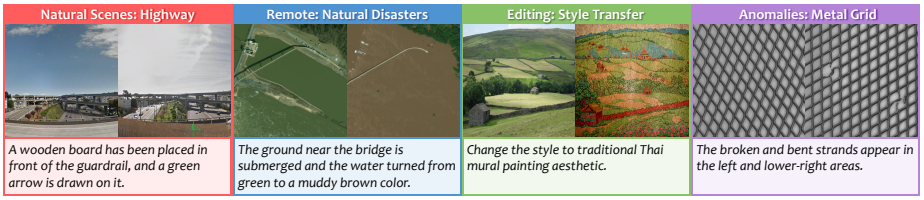

Fig. 3: Examples from C3-Bench. Each image pair is displayed with its Domain: Context, along with the corresponding human-annotated change description.

tasks across multiple communities, including change detection [2, 23, 67, 68, 82], image editing [13, 43, 65, 75, 108], and anomaly detection [10, 30, 107, 119]. By distilling the recurring patterns of change across these communities, we conceptualize 51 real-world change contexts, organized under four overarching domains.

Context Formalization. Building upon this context set, we formalize each context by defining context-specific change criteria $\mathcal{C}_c$ using a unified textual template (see Tab. 2). Each set of criteria comprises three components: *(i)* change to detect, specifying the target changes to be described; *(ii)* change to ignore, outlining differences that are irrelevant or non-essential; and *(iii)* tone and nuance, prescribing the desired linguistic style for expressing the detected changes [87]. This process totals 51 context-criteria pairs that delineate the semantic boundaries of each change—facilitating systematic and consistent evaluation of change descriptions within a shared, well-defined semantic space.

Data Collection and Pairing. We collect our change captioning data from more than fifteen publicly available benchmarks—including MVTecAD [10], SECOND [94], MagicBrush [108], PASLCD [29], and ChangeVPR [48]. Unlike previous evaluation datasets, we incorporate heterogeneous and open-domain visual sources, substantially broadening data diversity [117]. For crucial contexts not sufficiently covered by existing benchmarks, we supplement coverage by referring to publicly available imagery platforms such as Google Earth and Google Street View. We construct high-fidelity image pairs from existing correspondences or via feature matching [58] and UTM-guided sampling [11, 48], followed by careful

manual verification. Lastly, each pair is assigned to its corresponding context set and forwarded to the annotation stage.

Data Annotation. To ensure high-quality annotations, all annotators are provided with auxiliary metadata (e.g., change masks [2, 48, 108], building damage polygons [35], and anomaly categories [10, 107, 119]) overlaid on or displayed alongside the original images. Clear context-specific criteria are also provided to standardize the textual definition of relevant changes. All annotations are then converted into a unified schema for systematic downstream analysis.

Data Filtering and Quality Control. Each image pair is jointly reviewed by annotators and domain experts to ensure context-level consistency and semantic coherence. The candidate pairs then undergo a multi-stage validation pipeline—involving redundancy filtering, temporal verification, and caption-image consistency checks, followed by context-wise rebalancing to meet the target data size. Fig. 3 illustrates several examples from the curated C3-Bench. A full list of contexts, data sources, auxiliary metadata, as well as further details and examples, is provided in the Appendix.

4.2 Evaluation Metrics

To facilitate more human-aligned evaluation, we adopt the LLM-as-Judge framework [32,114]—to the best of our knowledge, for the first time in change captioning. We instruct GPT-5.2 [57] to assess each model-generated caption against the ground-truth description across four dimensions: correctness, specificity, fluency, and relevance. The *Correctness* metric refers to how closely the model prediction aligns with the ground-truth [53, 83]. The *specificity* measures whether the predicted description provides a concrete explanation rather than a generic statement [41]. The *fluency* metric evaluates the tone and naturalness of model prediction [66, 87], whereas the *relevance* metric determines whether model prediction provides answers directly related to the ground-truth [14]. We also prompt GPT-5.2 to provide a single *aggregation* score for a rapid, end-to-end comparison [117]. To ensure fair evaluation of leveraging GPT-5.2 as a judge, we display consistent results using Claude-Haiku-4.5 [5] in the Appendix.

Beyond these metrics, we also evaluate *reversibility*, grounded in the premise that a reliable change understanding should remain semantically consistent under input-order reversal—differing only in directionality rather than exhibiting order-induced artifacts [48,76]. Specifically, given a change description from $\mathcal{I}_A$ to $\mathcal{I}_B$ and its counterpart $\mathcal{I}_B$ to $\mathcal{I}_A$, we instruct GPT-5.2 to determine whether the two captions describe the same underlying changes with opposite directionality—assigning a score of 1 if reversible and 0 otherwise. This metric is partially related to position bias [76] and temporal consistency [48,86], in that they all explore the robustness to input image ordering. However, fundamental differences are that we investigate reversibility at the level of open-ended descriptions and evaluate order effects on single image pairs rather than on long image sequences [18,76,93]. All evaluation prompts are provided in the Appendix.

Table 3: System prompt for context-aware change captioning for LMMs. The varying parts are marked in yellow. During inference, these parts are dynamically replaced with the context-specific change criteria (Tab. 2) and associated image pairs.

Segment	Description
Task	Analyze an image pair and describe the changes from the <before> image to the <after> image according to the given change criteria.
Change Criteria	<context-specific change criteria>
Output Format	Produce a single coherent paragraph describing only the detected changes. Enclose the final answer within <output> ... </output> tags.
Image Pair	Each image is preceded by a corresponding tag. <before> Image $\mathcal{I}_A$ <after> Image $\mathcal{I}_B$

5 Experiments and Results

5.1 Overall Setups

We benchmark 32 models on C3-Bench—including 6 SOTA conventional change captioning models, 9 leading proprietary LMMs, and 17 open-source LMMs. The conventional models include DUDA [60], CLIP4IDC [34], VARD [77], SCORER [80], and DIRL [79]. The proprietary LMMs involve GPT-5.2, GPT-5.1, GPT-5-mini, GPT-4o, and GPT-4o-mini [1,57], Gemini-3-Pro, Gemini-3-Flash, Gemini-2.5-Pro, and Gemini-2.5-Flash [31]. The open-source LMMs comprise Llama-3.2-Vision-Instruct (11B / 90B) and Llama-4.0-Scout-17B-16E [25], Qwen2.5-VL (3B / 7B / 32B / 72B) [7] and Qwen3-VL (2B / 8B / 32B) [6], InternVL3 (2B / 8B / 38B / 78B) [118] and InternVL3.5 (2B / 8B / 38B) [88]. For consistency, all LMMs employ a temperature of 0 and a maximum completion length of 2048 tokens. Images are standardized to 512×512 pixel resolution. We adopt correctness, specificity, fluency, relevance, aggregation (all 1–10 scale), and reversibility rate (0–1) as our primary metrics. Since all changes are inherently reversible [48], the reversibility rate yields an upper bound of 1.

5.2 Evaluation Protocols

Humans. We sample a subset of 400 image pairs—balanced across domains and contexts—which we refer to as C3-Bench-small. Human evaluators independently describe the changes in each image pair with the corresponding criteria, and their descriptions are scored using the aforementioned metrics. For comparison, we also report the performance of GPT-5.2 on C3-Bench-small. Further details of the human evaluators' setups are provided in the Appendix.

Conventional Models. We train each conventional change captioning model on four representative datasets: CLEVR-CC [60], LEVIR-CC [52], SpotTheDiff [45], and IER [75]. This training process yields four distinctive models for each method. Then, we assess each of the four distinctive models on C3-Bench; this process yields 51 assessments per model, given the 51 context-criteria pairs. As C3-Bench subsumes a broad spectrum of real-world change contexts (including all real-world contexts covered by aforementioned datasets) with diverse data

Table 4: User study results. We measure Pearson (r), Spearman (ρ), and Kendall (τ) correlation coefficients to quantify alignment with human judgments.

(a) Scoring-based judge.

Metric	r	ρ	τ
BLEU-4 [59]	0.3314	0.3202	0.2233
ROUGE-L [51]	0.4916	0.4460	0.3112
chrF [63]	0.4548	0.4174	0.2879
BERTScore [109]	0.4494	0.4443	0.3161
CLIP-S [37]	0.2542	0.2931	0.2063
Aggregation	**0.7563**	**0.7944**	**0.6317**
Correctness	0.7173	0.7448	0.5963
Specificity	0.7268	0.7378	0.5812
Fluency	0.6193	0.5839	0.4857
Relevance	0.6382	0.6427	0.4985

(b) Pair-based judge.

Metric	r	ρ	τ
BLEU-4 [59]	0.0146	-0.0026	-0.0021
ROUGE-L [51]	0.1129	0.1073	0.0880
chrF [63]	0.1610	0.1596	0.1308
BERTScore [109]	0.1017	0.0948	0.0777
CLIP-S [37]	-0.0107	-0.0028	-0.0023
Aggregation	**0.6118**	**0.6107**	**0.5156**

(c) Reversibility judge.

Metric	r	ρ	τ
Reversibility	**0.8674**	**0.7670**	**0.7025**

sources, the proposed protocol investigates the potential generalizability of the implicit semantics inherited in θ beyond data-specific performances.

LMMs. Depart from conventional models, LMMs can be explicitly conditioned on $\mathcal{C}_c$ through textual instructions. To facilitate fair and reproducible comparisons, we leverage all LMMs in an off-the-shelf setting, without any task-specific fine-tuning. We also avoid advanced prompt strategies [89, 90] and multimodal prompting methods [8, 33]. Instead, all LMMs are evaluated on C3-Bench using a structured prompt template shown in Tab. 3. Here, a structured system prompt is combined with $\mathcal{C}_c$ in a plug-and-play manner for each of the 51 context-criteria pairs—explicitly capturing the contextual semantics. We also tag each input image pair to indicate its temporal order within the before–after sequence. During evaluation, the performance of $\hat{y}^c_{A\to B}$ is compared to the ground-truth caption $y^c_{A\to B}$ for each shared context c. Reversibility is evaluated by comparing $\hat{y}^c_{A\to B}$ against $\hat{y}^c_{B\to A}$ (for conventional models, $\hat{y}_{A\to B}$ against $\hat{y}_{B\to A}$). This protocol parallels the context-wise evaluation used for conventional models, enabling unified comparison across model families while systematically investigating the capabilities of LMMs in a more principled and consistent manner.

5.3 Alignment with Human Judgments

Before benchmarking performance on C3-Bench, we validate our proposed evaluation metrics with human judgments through a user study involving 40 participants on C3-Bench-small (see Tab. 4). We sample 2,800 change descriptions generated by seven LMMs (the latest models from each generation), and split them into 1,400 instances for scoring-based and 1,400 for pair-based protocols [16]; additional details on the user study are provided in the Appendix. The results show that all proposed metrics exhibit consistently stronger alignment with human judgments than conventional metrics across both protocols. Surprisingly, GPT-5.2 achieves a strong correlation with human judgments on reversibility (Pearson $r > 0.80$)—indicating high fidelity as a reversibility judge and consistent with prior observations that large language models can approx-

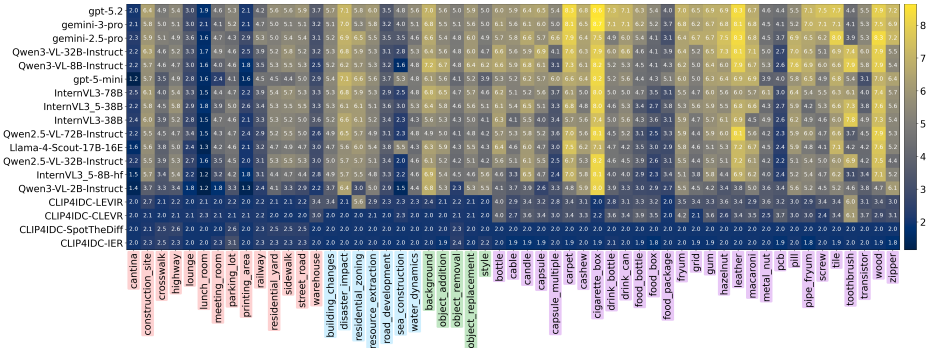

Fig. 4: Fine-grained results across domains and contexts on C3-Bench. The performance is represented as an Aggregation score, where each context is highlighted with its corresponding domain: Natural Scenes, Remote Sensing, Image Editing, and Anomalies. The actual score of each model on a given context is shown in the corresponding cell, with higher scores indicated by higher color intensity.

imate human judgments [16, 26, 27, 55, 112, 114]. Importantly, beyond improved human alignment, our metrics uniquely enable fine-grained analysis of individual change descriptions while remaining fully compatible with rapid, end-to-end comparisons. Collectively, these contributions establish a human-aligned foundation within our C3-Bench framework that reveals genuine yet previously obscured real-world generalization of contemporary change captioning approaches.

5.4 Benchmark Results

Tab. 5 summarizes the results of 32 models on C3-Bench. We illustrate our key observations as follows:

Human Level Performance. Not surprisingly, human evaluators demonstrate high performance across all semantic dimensions—outperforming the best LMM (GPT-5.2) by 1.73 points in Aggregation. Furthermore, human performance on the reversibility rate is remarkably high (0.93)—indicating a strong intuitiveness in understanding symmetrical changes. In contrast, the gap narrows considerably on Fluency, suggesting LMM-generated sentences can be comparable or even superior in linguistic coherence and naturalness—demonstrating their effectiveness in producing high-quality textual content [117].

Limited Generalization of Conventional Models. Across C3-Bench, conventional methods exhibit consistently low performance, while maintaining relatively high Fluency metric (7.73–8.29). Notably, this performance collapse persists even for recent state-of-the-art models such as DIRL [79], which have reported strong results on previous evaluation benchmarks. These findings underscore two broader implications: *(i)* while conventional models can generate fluent descriptions, they fail to capture the intended change semantics under diverse real-world change contexts; and *(ii)* existing evaluation benchmarks, largely op-

Table 5: C3-Bench results. Mean and standard deviation are reported over three GPT-5.2 runs. † denotes results on C3-Bench-small. Conventional models are averaged over trained weights. Human scores are in violet; best model scores are in **bold**.

Method	Correctness	Specificity	Fluency	Relevance	Aggregation	Reversibility
Conventional Models						
DUDA [60]	1.39 ± 0.01	1.43 ± 0.05	7.81 ± 0.25	2.11 ± 0.23	2.09 ± 0.09	0.15 ± 0.00
R^3Net [81]	1.39 ± 0.01	1.61 ± 0.09	8.26 ± 0.11	2.67 ± 0.19	2.39 ± 0.06	0.13 ± 0.00
CLIP4IDC [34]	1.40 ± 0.02	1.78 ± 0.05	8.07 ± 0.13	2.77 ± 0.20	2.43 ± 0.06	0.21 ± 0.00
SCORER [80]	1.39 ± 0.01	1.56 ± 0.07	7.98 ± 0.12	2.74 ± 0.20	2.36 ± 0.06	0.18 ± 0.00
VARD [77]	1.44 ± 0.02	1.56 ± 0.06	8.29 ± 0.12	2.75 ± 0.20	2.40 ± 0.07	0.28 ± 0.00
DIRL [79]	1.36 ± 0.01	1.53 ± 0.05	7.73 ± 0.15	2.71 ± 0.22	2.30 ± 0.04	0.20 ± 0.00
C3-Bench-small Performance						
Human Level†	6.96 ± 0.10	7.12 ± 0.08	8.32 ± 0.23	7.89 ± 0.15	7.45 ± 0.18	0.93 ± 0.00
GPT-5.2† [57]	5.64 ± 0.12	5.27 ± 0.17	8.53 ± 0.25	6.33 ± 0.17	5.72 ± 0.20	0.61 ± 0.00
Proprietary LMMs						
GPT-4o-mini [1]	4.56 ± 0.09	4.16 ± 0.13	8.67 ± 0.23	5.23 ± 0.16	4.65 ± 0.17	0.29 ± 0.00
GPT-4o [1]	4.95 ± 0.10	4.34 ± 0.15	8.48 ± 0.36	5.83 ± 0.26	4.95 ± 0.19	0.49 ± 0.00
GPT-5-mini [57]	4.87 ± 0.13	4.60 ± 0.21	8.34 ± 0.22	5.00 ± 0.16	4.87 ± 0.19	0.45 ± 0.00
GPT-5.1 [57]	5.43 ± 0.13	5.12 ± 0.20	8.61 ± 0.29	5.98 ± 0.17	5.49 ± 0.20	0.63 ± 0.00
GPT-5.2 [57]	$\mathbf{5.47 \pm 0.11}$	$\mathbf{5.16 \pm 0.18}$	8.65 ± 0.30	$\mathbf{5.98 \pm 0.13}$	$\mathbf{5.51 \pm 0.20}$	0.62 ± 0.00
Gemini-2.5-Flash [31]	4.99 ± 0.13	4.69 ± 0.17	8.26 ± 0.33	5.69 ± 0.15	5.04 ± 0.20	0.48 ± 0.00
Gemini-2.5-Pro [31]	5.18 ± 0.10	4.92 ± 0.15	8.37 ± 0.32	5.85 ± 0.12	5.23 ± 0.18	0.52 ± 0.00
Gemini-3-Flash [31]	5.42 ± 0.10	5.08 ± 0.14	8.34 ± 0.30	5.86 ± 0.11	5.39 ± 0.17	0.62 ± 0.00
Gemini-3-Pro [31]	5.45 ± 0.10	5.15 ± 0.16	8.40 ± 0.25	5.95 ± 0.17	5.50 ± 0.19	$\mathbf{0.73 \pm 0.00}$
Open-source LMMs						
Llama-3.2-11B-Vision-Ins. [25]	2.45 ± 0.06	2.51 ± 0.12	8.14 ± 0.29	2.89 ± 0.13	2.56 ± 0.12	0.12 ± 0.00
Llama-3.2-90B-Vision-Ins. [25]	2.55 ± 0.10	2.62 ± 0.13	8.20 ± 0.20	2.93 ± 0.11	2.65 ± 0.14	0.20 ± 0.00
Llama-4-Scout-17B-16E-Ins. [25]	4.63 ± 0.09	4.43 ± 0.13	8.46 ± 0.29	5.19 ± 0.23	4.71 ± 0.17	0.44 ± 0.00
Qwen2.5-VL-3B-Ins. [7]	2.87 ± 0.04	2.75 ± 0.07	8.21 ± 0.21	3.90 ± 0.21	3.00 ± 0.15	0.60 ± 0.00
Qwen2.5-VL-7B-Ins. [7]	4.26 ± 0.09	3.99 ± 0.14	8.63 ± 0.27	5.07 ± 0.21	4.41 ± 0.18	0.31 ± 0.00
Qwen2.5-VL-32B-Ins. [7]	4.51 ± 0.11	4.28 ± 0.16	8.30 ± 0.23	5.06 ± 0.25	4.57 ± 0.20	0.40 ± 0.00
Qwen2.5-VL-72B-Ins. [7]	4.54 ± 0.11	4.38 ± 0.17	8.36 ± 0.29	5.06 ± 0.23	4.61 ± 0.19	0.38 ± 0.00
Qwen3-VL-2B-Ins. [6]	3.84 ± 0.10	3.74 ± 0.16	8.56 ± 0.33	4.38 ± 0.24	3.92 ± 0.19	0.28 ± 0.00
Qwen3-VL-8B-Ins. [6]	5.07 ± 0.12	4.76 ± 0.18	8.68 ± 0.32	5.54 ± 0.13	5.10 ± 0.21	0.40 ± 0.00
Qwen3-VL-32B-Ins. [6]	5.18 ± 0.14	4.93 ± 0.22	8.49 ± 0.32	5.58 ± 0.18	5.16 ± 0.23	0.47 ± 0.00
InternVL3-2B [118]	3.19 ± 0.06	2.96 ± 0.10	8.72 ± 0.21	4.08 ± 0.26	3.35 ± 0.13	0.28 ± 0.00
InternVL3-8B [118]	3.93 ± 0.07	3.68 ± 0.10	8.78 ± 0.22	4.77 ± 0.22	4.10 ± 0.16	0.36 ± 0.00
InternVL3-38B [118]	4.72 ± 0.09	4.34 ± 0.13	8.16 ± 0.19	5.46 ± 0.29	4.72 ± 0.19	0.46 ± 0.00
InternVL3-78B [118]	4.86 ± 0.09	4.51 ± 0.14	8.66 ± 0.28	5.59 ± 0.24	4.94 ± 0.19	0.45 ± 0.00
InternVL3.5-2B [88]	3.20 ± 0.06	3.08 ± 0.10	$\mathbf{8.94 \pm 0.23}$	4.32 ± 0.20	3.41 ± 0.18	0.44 ± 0.00
InternVL3.5-8B [88]	4.30 ± 0.08	4.16 ± 0.13	8.63 ± 0.23	4.92 ± 0.21	4.44 ± 0.16	0.33 ± 0.00
InternVL3.5-38B [88]	4.83 ± 0.08	4.52 ± 0.13	8.46 ± 0.28	5.54 ± 0.24	4.92 ± 0.18	0.48 ± 0.00

timized for fixed reference matching with limited contextual diversity, may not stress-test a model's ability to adapt to diverse real-world changes.

LMM Landscapes in Change Captioning. In general, proprietary models achieve higher Aggregation scores, ranging from 4.65 to 5.51, led by GPT-5.2. For open-source models, small variants (e.g., Llama-3.2-Vision-11B or Qwen2.5-VL-3B) surpass conventional models but remain limited in semantic understanding. Qwen3-VL-32B delivers strong overall performance, yielding an Aggregation of 5.16 ± 0.23 that matches Gemini-2.5-Pro (5.23 ± 0.18) and trails GPT-5.2 by only 0.35 points. These findings indicate that open-source LMMs have almost bridged the gap with proprietary models, with remaining challenges mainly arising from semantic accuracy rather than linguistic coherence. Overall, the results exhibit a monotonic gain as LMMs scale in size and advance across generations, consistent with observations reported in recent multimodal benchmarks [22, 100, 101, 105].

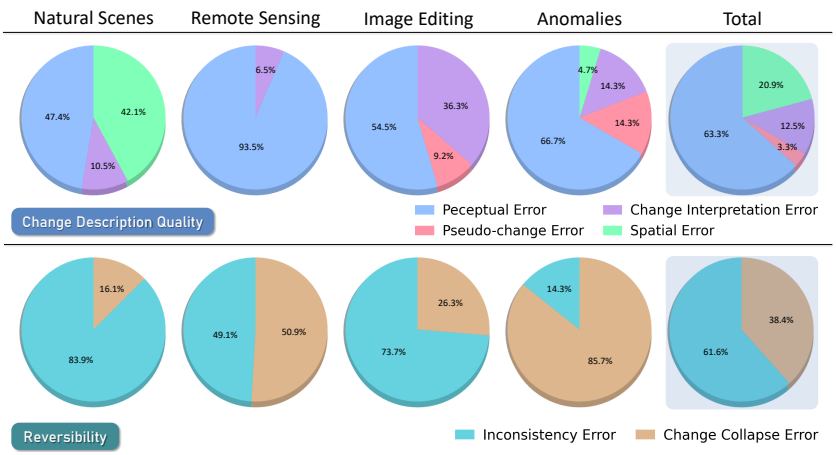

Fig. 5: Error distributions across domains. *(i)* change description quality errors (Top); and *(ii)* reversibility errors (Bottom).

Fine-grained Results across Domains and Contexts. As shown in Fig. 4, several models, including GPT-5-mini and Llama-4-Scout-17B-16E, exhibit pronounced strengths on specific contexts such as sea constructions [71] (in Remote Sensing), indicating that their gains are often highly localized rather than uniformly distributed across contexts. At the domain level, open-source models such as Qwen3-VL-8B surpass proprietary models in the Image Editing domain. Conventional models peak on only a few contexts and perform poorly elsewhere, due to architectural constraints that limit them to fixed, training-time semantics; further discussion on this and directions for future work are in the Appendix.

Do Models Understand Change Symmetrically? Notably, we discover that conventional models exhibit limited reversibility, revealing severe order sensitivity and a failure to form stable, order-invariant change representations. While LMMs show improved robustness, significant order bias remains: Gemini-3-Pro achieves the highest reversibility of 0.73; yet, open-source LMMs exhibit stronger sensitivity to image ordering—hindering a reliable understanding of changes [48]. This highlights that position bias [76, 93] in LMMs is not confined to long image sequences (e.g., 5–100 images) but is already pronounced in simple image pairs.

5.5 Error Analysis

To quantify and characterize the main bottlenecks of the best-performing model on our benchmark, we investigate its errors on C3-Bench-small. We first categorize failures related to change description quality into four distinct types and provide a clear definition upon examination:

1. **Perceptual Error**, arising from unrecognized objects, misclassified object categories, or incorrect object attributes.
2. **Pseudo-change Error**, describing context-irrelevant objects or states as meaningful changes.

Table 6: Ablation study. The default configuration, marked in violet, corresponds to the setting in Tab. 5.

(a) Ablation on the prompt composition.

No.	Prompt composition			Metrics					
	Criteria	Sequential tags	Reasoning order	Correctness	Specificity	Fluency	Relevance	Aggregation	Reversibility
1		✓	forward	4.83 ± 0.10	4.58 ± 0.14	8.60 ± 0.27	5.32 ± 0.16	4.93 ± 0.18	0.58 ± 0.00
2	✓		forward	5.27 ± 0.06	5.16 ± 0.12	8.63 ± 0.28	5.73 ± 0.12	5.35 ± 0.09	0.60 ± 0.00
3	✓	✓	backward	5.27 ± 0.09	4.94 ± 0.16	8.60 ± 0.27	5.81 ± 0.11	5.31 ± 0.18	0.62 ± 0.00
4	✓	✓	forward	**5.47 ± 0.11**	**5.16 ± 0.18**	**8.65 ± 0.30**	**5.98 ± 0.13**	**5.51 ± 0.20**	**0.62 ± 0.00**

(b) Ablation on the criteria composition.

No.	Criteria composition			Metrics					
	Ch. to Detect	Ch. to Ignore	Tone & Nuance	Correctness	Specificity	Fluency	Relevance	Aggregation	Reversibility
1		✓	✓	5.17 ± 0.10	4.91 ± 0.15	8.61 ± 0.24	5.63 ± 0.16	5.12 ± 0.13	0.60 ± 0.00
2	✓		✓	5.32 ± 0.09	5.05 ± 0.14	8.52 ± 0.23	5.85 ± 0.16	5.44 ± 0.16	0.61 ± 0.00
3	✓	✓		5.47 ± 0.13	5.16 ± 0.19	8.17 ± 0.29	5.98 ± 0.16	5.49 ± 0.19	0.60 ± 0.00
4	✓	✓	✓	**5.47 ± 0.11**	**5.16 ± 0.18**	**8.65 ± 0.30**	**5.98 ± 0.13**	**5.51 ± 0.20**	**0.62 ± 0.00**

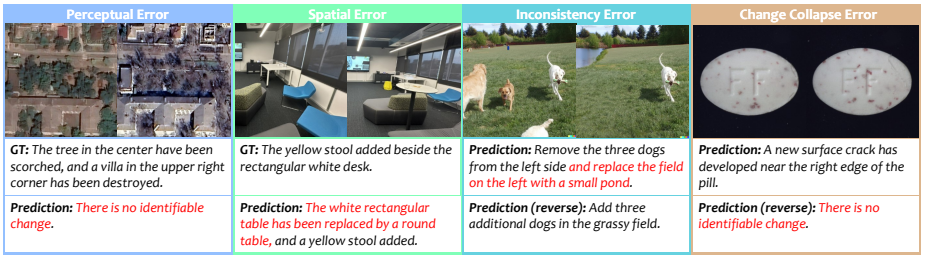

Fig. 6: Representative failure examples (Top-2 errors) of each error case in C3-Bench. Erroneous regions are highlighted in red.

3. **Change Interpretation Error**, failing to correctly infer how it has changed between the given image pair.
4. **Spatial Error**, misinterpreting the spatial relationships or locations associated with the changes.

Further, we define two reversibility-related errors as follows:

1. **Inconsistency Error**, producing semantically inconsistent descriptions that are not mutually inverse with omitted or extraneous key details.
2. **Change Collapse Error**, identifying a change in one direction but failing to detect a change at all in the reversed order.

As shown in Fig. 5, perceptual error emerges as the primary bottleneck (63.3% in total)—underscoring the significance of robust visual capabilities for accurate change captioning. The secondary bottleneck exhibits strong domain dependency, revealing distinct challenges across different domains. For instance, in the case of the Natural Scenes domain, where image pairs are often captured by moving vehicles or mobile robots [49], scene misalignment leads to a marked increase in spatial errors, causing models to confuse viewpoint-induced differences with genuine changes observed under a shared viewpoint. For reversibility-related errors, the inconsistency error is the most frequent type. However, within the Anomalies domain, change collapse errors dominate, where models fail to revert anomalous regions back to an intact, status quo state—even when the criteria explicitly require such recovery. This may suggest that the model's implicit bias

Table 7: Results of multi-dataset training. We use CLIP4IDC [34], the strongest conventional baseline. See Sec. 5.2 for datasets included in ALL.

No.	Training	Metrics				
		Correctness	Specificity	Fluency	Relevance	Aggregation
1	Single (LEVIR-CC [52])	1.78 ± 0.02	1.63 ± 0.03	8.81 ± 0.23	2.35 ± 0.22	1.79 ± 0.03
2	ALL [45, 52, 60, 75]	1.31 ± 0.01	1.24 ± 0.02	4.21 ± 0.34	1.58 ± 0.09	1.31 ± 0.02

in anomalous image pairs predisposes it to focus on abnormal cues rather than restoring the status quo—highlighting that a misalignment between biases and the criteria can substantially affect the performance. Fig. 6 illustrates representative failure cases; more examples and analysis are provided in the Appendix.
Why Models Cannot? Reversibility errors may arise from models being unidirectionally trained on sequentially ordered image-text data and from the causal attention mechanism in their forward process. Description errors may stem from models being optimized for image-text alignment rather than for fine-grained perception or explicit geometry—lacking in visual and spatial understanding.

5.6 Ablation Study on LMM Prompting

To gain further insight into LMM-based change captioning, we ablate key prompt components—change criteria, sequential tags, and reasoning order (*before*-image-first vs. *after*-image-first). As illustrated in Tab. 6 (a), removing explicit criteria and relying on generic instructions leads to substantial degradation, indicating that model's implicit biases are misaligned with context-relevant semantics—underlining the importance of our criteria-conditioned prompt design over previous naive prompting schemes [4, 40, 110, 116]; while all components contribute meaningfully, Change to Detect plays a key role in enhancing contextual alignment (see Tab. 6(b)). Excluding image tags also deteriorates performance, underscoring the advantages of explicitly grounding sequential order for accurate and reversible change descriptions. Finally, backward reasoning, presenting the *after*-image first, further impairs performance, indicating that misalignment between image order and reasoning direction induces additional position bias. This finding suggests that LMMs are biased to canonical temporal ordering (*before*-first), likely due to training-time exposure to temporally ordered image-text data.

5.7 Multi-Dataset Training for Change Captioning

Can training on multiple datasets concurrently improve change captioning performance, as in other data-driven computer vision tasks [44, 61, 95, 96]? As shown in Tab. 7, simple unification does not lead to a performance gain—even worsening the results. This can be attributed to two main factors: *(i)* the size of the unified dataset is still insufficient to induce a discriminative latent space to bridge the semantic gap across heterogeneous change contexts, and *(ii)* contextual conflicts between the implicit criteria of each dataset (e.g., color differences are treated as meaningful in CLEVR-CC [60], whereas they are regarded as pseudo-changes in LEVIR-CC [52]; and IER [75] requires an instructional tone whereas most other

datasets adopt a descriptive tone) confuse the overall learning signal, highlighting the significance of our context-aware formulation for coherent data scaling.

6 Conclusion

In this paper, we introduced C3-Bench, a comprehensive benchmark for evaluating Context-aware Change Captioning. Based on a distinct problem formulation, C3-Bench provides a dataset with unprecedented diversity spanning multiple domains and contexts, along with principled metrics, establishing a robust testbed for both conventional models and modern LMMs. Through extensive experiments, we revealed key challenges of existing benchmarks and model families, including the limited generalizability of conventional models, critical bottlenecks in LMMs, pronounced pairwise position bias, and crucial components in prompt design. We expect C3-Bench to lay a solid foundation and guide future research toward more robust and reliable change understanding.

Acknowledgements

This research was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220926, Development of Self-directed Visual Intelligence Technology Based on Problem Hypothesis and Self-supervised Methods); by the InnoCORE program of the Ministry of Science and ICT (GIST InnoCORE KH0860); by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009989).

References

1. Achiam, O.J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., ing Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., laine Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., abella Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick,

L., Kim, J.W., Kim, C., Kim, Y., Kirchner, H., Kiros, J.R., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C., Lim, R., Lin, M., Lin, S., teusz Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., drey Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D.P., Mu, T., Murati, M., Murk, O., M'ely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Long, O., O'Keefe, C., Pachocki, J.W., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Pokorny, M., Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J.W., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M.D., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N.A., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C.L., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., ing Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2023)

2. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Autonomous Robots **42**, 1301 – 1322 (2016)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. ArXiv **abs/1607.08822** (2016)
4. Anonymous: Imagine how to change: Explicit procedure modeling for change captioning. In: The Fourteenth International Conference on Learning Representations (2026)
5. Anthropic: Claude developer platform (nd), `https://claude.com/platform/api`, accessed: 2026-01-22
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L.Y., Ren, X., yi Ren, X., Song, S., Sun, Y.C., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J.X., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025)
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025)
8. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? 2024 IEEE/CVF Conference on Computer

Vision and Pattern Recognition Workshops (CVPRW) pp. 1539–1550 (2024)
9. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: IEEvaluation@ACL (2005)
10. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9584–9592 (2019)
11. Berton, G.M., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4868–4878 (2022)
12. Black, A., Shi, J., Fai, Y., Bui, T., Collomosse, J.P.: Vixen: Visual text comparison network for image difference captioning **abs/2402.19119** (2024)
13. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18392–18402 (2022)
14. Chakraborty, S., Ghosal, S.S., Yin, M., Manocha, D., Wang, M., Bedi, A.S., Huang, F.: Transfer q star: Principled decoding for llm alignment (2024)
15. Chang, S., Ghamisi, P.: Changes to captions: An attentive network for remote sensing change captioning. IEEE Transactions on Image Processing **32**, 6047–6060 (2023)
16. Chen, D., Chen, R., Zhang, S., Liu, Y., Wang, Y., Zhou, H., Zhang, Q., Zhou, P., Wan, Y., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: International Conference on Machine Learning (2024)
17. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: Lion : Empowering multimodal large language model with dual-level visual knowledge. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 26530–26540 (2023)
18. Chen, J., Xu, D., Fei, J., Feng, C.M., Elhoseiny, M.: Document haystacks: Vision-language reasoning over piles of 1000+ documents. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 24817–24826 (2024)
19. Chen, K.J., Li, W., Lei, S., Chen, J., Jiang, X., Zou, Z., Shi, Z.X.: Continuous remote sensing image super-resolution based on context interaction in implicit function space. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023)
20. Chen, K., Li, W., Chen, J., Zou, Z., Shi, Z.X.: Resolution-agnostic remote sensing scene classification with implicit neural representations. IEEE Geoscience and Remote Sensing Letters **20**, 1–5 (2023)
21. Chen, K., Liu, C., Li, W., Liu, Z., Chen, H., Zhang, H., Zou, Z., Shi, Z.X.: Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium pp. 8581–8584 (2023)
22. Chen, L., Li, J., wen Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? ArXiv **abs/2403.20330** (2024)
23. Chen, S., Yang, K., Stiefelhagen, R.: Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. 2021 IEEE Intelligent Vehicles Symposium (IV) pp. 502–509 (2021)
24. Du, P., Liu, P., Xia, J., Feng, L., Liu, S., Tan, K., Cheng, L.: Remote sensing image interpretation for urban environment analysis: Methods, system and examples. Remote. Sens. **6**, 9458–9474 (2014)

25. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A.S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., tian Cantón Ferrer, C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E.A., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I.M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J.Q., Alwala, K.V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K.R., El-Arini, K., Iyer, K., Malik, K., ley Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., hesh Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M.H.M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., lay Bashlykov, N., Bogoychev, N., Chatterji, N.S., Duchenne, O., cCelebi, O., Alrassy, P., Zhang, P., Li, P., Vasić, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., nie Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., hana Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S.C., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., ginie Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., ney Meers, W., Martinet, X., Wang, X., Tan, X.E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A.K., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, P.Y.B., Loyd, B., de Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, S.W., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzm'an, F., Kanayet, F.J., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G.G., Zhang,

G., Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., KamHou, U., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., Lavender, A., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., ish Bansal, M., Santhanam, N., Parks, N., White, N., ata Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N.P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., dro Rittner, P., Bontrager, P., Roux, P., Dollár, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Shankar, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S.K., Cho, S.B., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V.A., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Wang, Y., Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z.: The llama 3 herd of models (2024)

26. Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., Hashimoto, T.: Alpacafarm: A simulation framework for methods that learn from human feedback. ArXiv **abs/2305.14387** (2023)
27. Fu, J., Ng, S.K., Jiang, Z., Liu, P.: Gptscore: Evaluate as you desire. In: North American Chapter of the Association for Computational Linguistics (2023)
28. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. ArXiv **abs/2404.12390** (2024)
29. Galappaththige, C.J., Lai, J., Windrim, L., Dansereau, D.G., Suenderhauf, N., Miller, D.: Multi-view pose-agnostic change localization with zero labels. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11600–11610 (2025)
30. Gao, B.B.: Metauas: Universal anomaly segmentation with one-prompt meta-learning. ArXiv **abs/2505.09265** (2025)
31. Google: Google AI Studio Documentation. `https://aistudio.google.com/` (2025), accessed: 2025-12-12
32. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., Guo, J.: A survey on llm-as-a-judge. ArXiv **abs/2411.15594**

(2024)
33. Guo, Q., Mello, S.D., Yin, H., Byeon, W., Cheung, K.C., Yu, Y., Luo, P., Liu, S.: Regiongpt: Towards region understanding vision language model. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13796–13806 (2024)
34. Guo, Z., Wang, T., Laaksonen, J.T.: Clip4idc: Clip for image difference captioning. ArXiv **abs/2206.00629** (2022)
35. Gupta, R., Goodman, B., Patel, N.N., Hosfelt, R., Sajeev, S., Heim, E.T., Doshi, J., Lucas, K., Choset, H., Gaston, M.E.: xbd: A dataset for assessing building damage from satellite imagery. ArXiv **abs/1911.09296** (2019)
36. Han, L., Wong, D.F., Chao, L.S.: Lepor: A robust evaluation metric for machine translation with augmented factors. In: International Conference on Computational Linguistics (2012)
37. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. ArXiv **abs/2104.08718** (2021)
38. Hosseinzadeh, M., Wang, Y.: Image change captioning by learning from an auxiliary task. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2724–2733 (2021)
39. Hu, E., Guo, L., Yue, T., Zhao, Z., Xue, S., Liu, J.: Onediff: A generalist model for image difference captioning. ArXiv **abs/2407.05645** (2024)
40. Hu, J., Zhong, G., Yuan, J., Pan, W., Wang, X.: Mct-ccdiff: Context-aware contrastive diffusion model with mediator-bridging cross-modal transformer for image change captioning. IEEE Transactions on Image Processing **34**, 3294–3308 (2025)
41. Hu, X., Gao, M., Hu, S., Zhang, Y., Chen, Y., Xu, T., Wan, X.: Are llm-based evaluators confusing nlg quality criteria? ArXiv **abs/2402.12055** (2024)
42. Huang, Q., Liang, Y., Wei, J., Yi, C., Liang, H., fung Leung, H., Li, Q.: Image difference captioning with instance-level fine-grained feature representation. IEEE Transactions on Multimedia **24**, 2004–2017 (2022)
43. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Zhou, Y., Xie, C.: Hq-edit: A high-quality dataset for instruction-based image editing. ArXiv **abs/2404.09990** (2024)
44. Jeong, C., Bae, I., Park, J.H., Jeon, H.G.: Test-time prompt tuning for zero-shot depth completion. 2025 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9443–9454 (2025)
45. Jhamtani, H., Berg-Kirkpatrick, T.: Learning to describe differences between pairs of similar images. ArXiv **abs/1808.10584** (2018)
46. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1988–1997 (2016)
47. Kim, H., Kim, J., Lee, H., a Park, H., Kim, G.: Viewpoint-agnostic change captioning with cycle consistency. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2075–2084 (2021)
48. Kim, J., Kim, U.: Towards generalizable scene change detection. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
49. Li, D., Chen, L., Xu, C., Kong, H.: Umad: University of macau anomaly detection benchmark dataset. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 5836–5843 (2024)
50. Li, R., Li, L., Zhang, J., Zhao, Q., Wang, H., Yan, C.: Region-aware difference distilling with attribute-guided contrastive regularization for change captioning. In: AAAI Conference on Artificial Intelligence (2025)

51. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Annual Meeting of the Association for Computational Linguistics (2004)
52. Liu, C., Zhao, R., Chen, H., Zou, Z., Shi, Z.X.: Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–20 (2022)
53. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023)
54. Liu, M., Xu, Z., Lin, Z., Ashby, T., Rimchala, J., Zhang, J., Huang, L.: Holistic evaluation for interleaved text-and-image generation. In: Conference on Empirical Methods in Natural Language Processing (2024)
55. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: Conference on Empirical Methods in Natural Language Processing (2023)
56. Lv, F., Wang, R., Jing, L.: Revisiting change captioning from self-supervised global-part alignment. In: AAAI Conference on Artificial Intelligence (2025)
57. OpenAI: OpenAI Platform Documentation. `https://platform.openai.com/docs/overview` (2025), accessed: 2025-12-12
58. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023)
59. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (2002)
60. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4623–4632 (2019)
61. Park, J.M., Kim, U.H., Lee, S., Kim, J.H.: Dual task learning by leveraging both dense correspondence and mis-correspondence for robust change detection with imperfect matches. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13739–13749 (2022)
62. Park, K.R., Park, J., Kim, S.T., Lee, H.J., Kim, J.U.: Leveraging textual compositional reasoning for robust change captioning. ArXiv **abs/2511.22903** (2025)
63. Popovic, M.: chrf: character n-gram f-score for automatic mt evaluation. In: WMT@EMNLP (2015)
64. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1951–1960 (2021)
65. Ryu, S., Kim, K., Baek, E., Shin, D., Lee, J.: Towards scalable human-aligned benchmark for text-guided image editing. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18292–18301 (2025)
66. Sai, A.B., Mohankumar, A.K., Khapra, M.M.: A survey of evaluation metrics used for nlg systems. ACM Computing Surveys (CSUR) **55**, 1 – 39 (2020)
67. Sakurada, K., Shibuya, M., Wang, W.: Weakly supervised silhouette-based semantic scene change detection. 2020 IEEE International Conference on Robotics and Automation (ICRA) pp. 6861–6867 (2018)

68. Sakurada, K., Wang, W., Kawaguchi, N., Nakamura, R.: Dense optical flow based change detection network robust to difference of camera viewpoints. ArXiv **abs/1712.02941** (2017)
69. Sellam, T., Das, D., Parikh, A.P.: Bleurt: Learning robust metrics for text generation. In: Annual Meeting of the Association for Computational Linguistics (2020)
70. Shafique, A., Cao, G., Khan, Z., Asad, M., Aslam, M.: Deep learning-based change detection in remote sensing images: A review. Remote. Sens. **14**, 871 (2022)
71. Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L.: A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–16 (2021)
72. Shi, X., Yang, X., Gu, J., Joty, S.R., Cai, J.: Finding it at another side: A viewpoint-adapted matching encoder for change captioning. ArXiv **abs/2009.14352** (2020)
73. Snover, M.G., Dorr, B.J., Schwartz, R.M., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Conference of the Association for Machine Translation in the Americas (2006)
74. Sun, Y., Qiu, Y., Khan, M., Matsuzawa, F., Iwata, K.: The stvchrono dataset: Towards continuous change recognition in time. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 14111–14120 (2024)
75. Tan, H., Dernoncourt, F., Lin, Z.L., Bui, T., Bansal, M.: Expressing visual relationships via language. In: Annual Meeting of the Association for Computational Linguistics (2019)
76. Tian, X., Zou, S., Yang, Z., Zhang, J.: Identifying and mitigating position bias of multi-image vision-language models. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10599–10609 (2025)
77. Tu, Y., Li, L., Su, L., Du, J., Lu, K., Huang, Q.: Viewpoint-adaptive representation disentanglement network for change captioning. IEEE Transactions on Image Processing **32**, 2620–2635 (2023)
78. Tu, Y., Li, L., Su, L., Lu, K., Huang, Q.: Neighborhood contrastive transformer for change captioning. IEEE Transactions on Multimedia **25**, 9518–9529 (2023)
79. Tu, Y., Li, L., Su, L., Yan, C., Huang, Q.: Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In: European Conference on Computer Vision (2024)
80. Tu, Y., Li, L., Su, L., Zha, Z.J., Yan, C.C., Huang, Q.: Self-supervised cross-view representation reconstruction for change captioning. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2793–2803 (2023)
81. Tu, Y., Li, L., Yan, C.C., Gao, S., Yu, Z.: R^3net:relation-embedded representation reconstruction network for change captioning. In: Conference on Empirical Methods in Natural Language Processing (2021)
82. Varghese, A., Gubbi, J., Ramaswamy, A., Balamuralidhar, P.: Changenet: A deep learning architecture for visual change detection. In: ECCV Workshops (2018)
83. Vayani, A., Dissanayake, D., Watawana, H., Ahsan, N., Sasikumar, N., Thawakar, O., Ademtew, H.B., Hmaiti, Y., Kumar, A., Kukreja, K., Maslych, M., Al Ghallabi, W., Mihaylov, M.M., Qin, C., Shaker, A.M., Zhang, M., Ihsani, M.K., Esplana, A.G., Gokani, M., Mirkin, S., Singh, H., Srivastava, A., Hamerlik, E., Izzati, F.A., Maani, F.A., Cavada, S., Chim, J., Gupta, R., Manjunath, S., Zhumakhanova, K., Rabevohitra, F.H., Amirudin, A.H., Ridzuan, M., Kareem, D.N.A., More, K.P., Li, K., Shakya, P., Saad, M., Ghasemaghaei, A., Djanibekov, A., Azizov, D., Jankovic, B., Bhatia, N., Cabrera, A., Obando-Ceron, J., Otieno, O., Farestam, F., Rabbani, M., Ballah, S., Sanjeev, S., Shtanchaev, A., Fatima,

M., Nguyen, T., Kareem, A., Aremu, T., Xavier, N.A.Z., Bhatkal, A., Toyin, H.O., Chadha, A., Cholakkal, H., Anwer, R.M., Felsberg, M., Laaksonen, J., Solorio, T., Choudhury, M., Laptev, I., Shah, M., Khan, S., Khan, F.S.: All languages matter: Evaluating lmms on culturally diverse 100 languages. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19565–19575 (June 2025)
84. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4566–4575 (2014)
85. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., Yan, T., Mo, W.J., Liu, H.H., Lu, P., Li, C., Xiao, C., Chang, K.W., Roth, D., Zhang, S., Poon, H., Chen, M.: Muirbench: A comprehensive benchmark for robust multi-image understanding. ArXiv **abs/2406.09411** (2024)
86. Wang, G.H., Gao, B.B., Wang, C.: How to reduce change detection to semantic segmentation. Pattern Recognition **138**, 109384 (2023)
87. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., Zheng, F.: Caption anything: Interactive image description with diverse multimodal controls. ArXiv **abs/2305.02677** (2023)
88. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hao, H., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Gu, Y., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Zhou, B., Su, W., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. ArXiv **abs/2508.18265** (2025)
89. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E.H., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. ArXiv **abs/2203.11171** (2022)
90. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. ArXiv **abs/2201.11903** (2022)
91. Wei, S., Luo, Y., Wang, Y., Luo, C.: Robust multimodal learning via representation decoupling. In: European Conference on Computer Vision (2024)
92. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L.J., Lin, D., Wetzstein, G.: Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22227–22238 (2024)
93. Wu, T.H., Biamby, G., Quenum, J., Gupta, R., Gonzalez, J., Darrell, T., Chan, D.M.: Visual haystacks: Answering harder questions about sets of images. ArXiv **abs/2407.13766** (2024)
94. Yang, K., Xia, G.S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L.: Asymmetric siamese networks for semantic change detection in aerial images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–18 (2022)
95. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10371–10381 (2024)
96. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. ArXiv **abs/2406.09414** (2024)

97. Yang, S., Guo, S., Zhao, J., Furao, S.: Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study. Pattern Recognit. **148**, 110204 (2024)
98. Yao, L., Wang, W., Jin, Q.: Image difference captioning with pre-training and contrastive learning. In: AAAI Conference on Artificial Intelligence (2022)
99. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11** (2023)
100. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., Lei, J., Lu, Q., Chen, R., Xu, P., Zhang, R., Zhang, H., Gao, P., Wang, Y., Qiao, Y., Luo, P., Zhang, K., Shao, W.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. ArXiv **abs/2404.16006** (2024)
101. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. ArXiv **abs/2308.02490** (2023)
102. Yu, W., Zhang, X., Zhu, X.X., Gloaguen, R., Ghamisi, P.: Minenetcd: A benchmark for global mining change detection on remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
103. Yuan, W., Neubig, G., Liu, P.: Bartscore: Evaluating generated text as text generation. ArXiv **abs/2106.11520** (2021)
104. Yue, S., Tu, Y., Li, L., Yang, Y., Gao, S., Yu, Z.: I3n: Intra- and inter-representation interaction network for change captioning. IEEE Transactions on Multimedia **25**, 8828–8841 (2023)
105. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yin, M., Yu, B., Zhang, G., Sun, H., Su, Y., Chen, W., Neubig, G.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. ArXiv **abs/2409.02813** (2024)
106. Zhang, H., Chen, H., Zhou, C., Chen, K., Liu, C., Zou, Z., Shi, Z.X.: Bifa: Remote sensing image change detection with bitemporal feature alignment. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–17 (2024)
107. Zhang, J., Ding, R., Ban, M., Dai, L.: Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. IEEE Robotics and Automation Letters **9**, 2008–2015 (2023)
108. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. ArXiv **abs/2306.10012** (2023)
109. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. ArXiv **abs/1904.09675** (2019)
110. Zhang, X., Wen, H., Wu, J., Qin, P., Xue, H., Nie, L.: Differential-perceptive and retrieval-augmented mllm for change captioning. Proceedings of the 32nd ACM International Conference on Multimedia (2024)
111. Zhang, X., Yu, W., Pun, M.O., Shi, W.: Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. ISPRS Journal of Photogrammetry and Remote Sensing (2023)
112. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v(ision) as a generalist evaluator for vision-language tasks. ArXiv **abs/2311.01361** (2023)
113. Zhang, Y., Su, Y., Liu, Y., Wang, X., Burgess, J., Sui, E., Wang, C., Aklilu, J., Lozano, A., Wei, A., Schmidt, L., Yeung-Levy, S.: Automated generation of challenging multiple-choice questions for vision language model evaluation. 2025

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 29580–29590 (2025)
114. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. ArXiv **abs/2306.05685** (2023)
115. Zheng, Z., Ma, A., Zhang, L., Zhong, Y.: Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 15173–15182 (2021)
116. Zhong, G., Hu, J., Chen, J., Yuan, J., Pan, W.: Decider: Difference-aware contrastive diffusion model with adversarial perturbations for image change captioning. In: AAAI Conference on Artificial Intelligence (2025)
117. Zhou, P., Peng, X., Song, J., Li, C., Xu, Z., Yang, Y., Guo, Z., Zhang, H., Lin, Y., He, Y., Zhao, L., Liu, S., Li, T., Xie, Y., Chang, X., Qiao, Y., Shao, W., Zhang, K.: Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 56–66 (2024)
118. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025)
119. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision (2022)