

Leveraging Textual Compositional Reasoning for Robust Change Captioning

Kyu Ri Park^{1*}, Jiyoung Park^{1*}, Seong Tae Kim¹, Hong Joo Lee^{2†}, Jung Uk Kim^{1†}

¹Kyung Hee University, Yong-in, South Korea

²Technical University of Munich, Munich, Germany

{kyuri0924, jy0117, st.kim, ju.kim}@khu.ac.kr, hongjoo.lee@tum.de

Abstract

Change captioning aims to describe changes between a pair of images. However, existing works rely on visual features alone, which often fail to capture subtle but meaningful changes because they lack the ability to represent explicitly structured information such as object relationships and compositional semantics. To alleviate this, we present **CORTEX** (COMpositional Reasoning-aware TEXT-guided), a novel framework that integrates complementary textual cues to enhance change understanding. In addition to capturing cues from pixel-level differences, CORTEX utilizes scene-level textual knowledge provided by Vision Language Models (VLMs) to extract richer image text signals that reveal underlying compositional reasoning. CORTEX consists of three key modules: (i) an Image-level Change Detector that identifies low-level visual differences between paired images, (ii) a Reasoning-aware Text Extraction (RTE) module that use VLMs to generate compositional reasoning descriptions implicit in visual features, and (iii) an Image-Text Dual Alignment (ITDA) module that aligns visual and textual features for fine-grained relational reasoning. This enables CORTEX to reason over visual and textual features and capture changes that are otherwise ambiguous in visual features alone.

Introduction

Change captioning aims to generate natural language descriptions that explain the differences between two images captured at different time points (Jhamtani and Berg-Kirkpatrick 2018). Unlike traditional image captioning, which focuses on describing a single static image, change captioning requires models to detect and clearly describe meaningful differences between image pairs. This is especially important in applications where capturing fine-grained visual differences is critical, such as in surveillance (Hoxha et al. 2022) and medical imaging (Li et al. 2023).

Since it holds significant potential for various practical applications, extensive research has been conducted to develop algorithms that more effectively describe changes. SCORER (Tu et al. 2023b) introduces self-supervised cross-view contrastive alignment and cross-modal backward reasoning, which solely relies on visual features to distin-

*These authors contributed equally.

†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

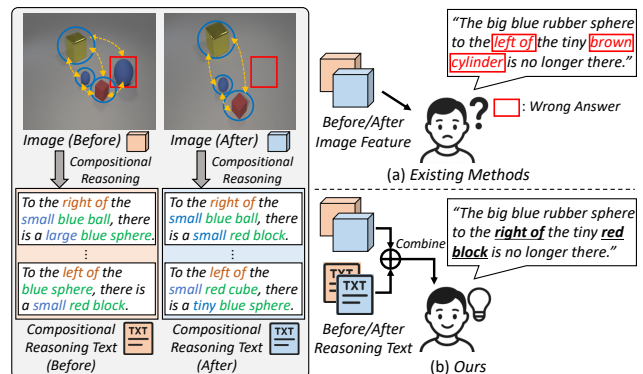


Figure 1: (a) Existing methods struggle to estimate changes because compositional reasoning cues are not explicitly represented in the image (*i.e.*, object relationships (yellow arrows), spatial arrangements (blue circles)). (b) In contrast, our method incorporates explicit textual compositional reasoning cues to enhance scene understanding, thereby enabling more accurate change description.

guish true semantic changes from pseudo changes caused by viewpoint shifts. SMART (Tu et al. 2024b) employs a multi-aspect relation learning network with a POS-based visual switch, relying on image data to capture semantic and positional relations. In addition, DURL (Tu et al. 2024a) uses self-supervised channel correlation and decorrelation to align captions with visual differences, effectively addressing pseudo changes from viewpoint and illumination variations.

Despite these advances, existing methods adopt a visual-only approach, relying on image-level features to describe changes. While these features can capture low-level appearance differences, they often fail to support *compositional reasoning*—the ability to understand structured semantics such as object relationships and spatial configurations. Since this type of information is not directly encoded in images but rather implicitly embedded (Hwang, Kim, and Kim 2023), models often struggle to generate accurate descriptions of changes. For example, as shown in Figure 1 (a), existing methods often misinterpret spatial relations (*‘left of’*) or misidentify reference objects (*‘a tiny brown cylinder’*). These limitations emphasize the need for a change caption-

ing model that can reason over compositional structures.

To address these limitations, we aim to enhance existing visual-only methods through a simple yet effective strategy that incorporates explicit textual cues conveying compositional reasoning. Unlike visual information, text can explicitly depict the structured semantics embedded in an image in a clear and interpretable form, serving as a strong signal for high-level reasoning (Hwang, Kim, and Kim 2023). As shown in Figure 1 (b), incorporating such compositional cues enables our model to better explain changes by capturing their relational and contextual meanings.

Based on the aforementioned points, we propose a novel **CO**mpositional **R**easoning-aware **TE**Xt-guided (CORTEX) framework for robust change captioning. It consists of three modules. (i) We employ an **image-level change detector** from visual-only approaches (Tu et al. 2024a, 2023b, 2024b) to identify visual differences between the input image pairs. (ii) To incorporate compositional understanding, we introduce a **Reasoning-aware Text Extraction (RTE) module**, which extracts explicit textual compositional reasoning cues (e.g., object relationships and spatial configurations). To this end, we leverage a widely adopted Vision-Language Model (VLM) with structured prompts to generate compositional descriptions for each image, offering rich semantic cues that enhance scene understanding. Also, (iii) we propose an **Image-Text Dual Alignment (ITDA) module** that aligns visual and textual features via static alignment (within the scene) and dynamic alignment (cross scene). This dual alignment allows the model to embed the textual compositional reasoning into the visual features while preserving their representational strengths, leading to richer scene representations and improved structural change reasoning.

As a result, CORTEX generates more accurate captions by effectively describing changes between two images. Notably, the core objective of this work is to overcome the limitations of visual-only change captioning methods by introducing two *plug-and-play* modules, namely RTE and ITDA, which can be seamlessly integrated into existing image-level change detectors to enhance compositional reasoning.

The major contributions of our paper are as follows:

- We devise CORTEX, a new plug-and-play framework that enhances the existing visual-only approaches by incorporating explicit textual compositional reasoning, which was previously embedded implicitly in images and challenging for models to accurately infer.
- We propose RTE module that leverages a VLM to extract structured textual cues that encode explicit compositional reasoning elements from images. The extracted text will be publicly released to facilitate future research.
- We introduce ITDA module, which aligns image and text features through static (within the scene) and dynamic (cross scene) alignment strategies to enable a more comprehensive understanding of both individual scenes and their differences.

Related Works

Change Captioning

Change captioning aims to generate natural language descriptions that capture semantic differences between two images, while filtering out irrelevant variations such as view-point or illumination changes. Earlier methods address irrelevant variations using techniques such as view-invariant representation learning, semantic alignment, and cross-modal regularization (Tu et al. 2023a,b, 2024b,a).

Recent methods aim to suppress irrelevant variations and enhance cross-modal alignment for more accurate change descriptions. RDD (Li et al. 2025) mitigates noise from global difference features and improves linguistic-visual consistency through region-aware difference distillation and attribute-guided contrastive learning. DECIDER (Zhong et al. 2025) addresses the limitations of auto-regressive decoders, such as error accumulation and weak inter-modality interaction. It adopts a contrastive diffusion framework with adversarial perturbations to generate more robust and semantically accurate change captions.

All visual-only methods struggle to capture fine-grained compositional and relational dynamics between objects. Our method addresses this gap by introducing textual compositionality into the training process, enabling more accurate and context-aware change descriptions.

Vision-Language Models in Vision Tasks

Vision-Language Models (VLMs) integrate visual and textual data to enable comprehensive image-text understanding (Zhang et al. 2024a). As a result, they have been successfully applied to various vision tasks—including image classification and retrieval (Radford et al. 2021), VQA (Khan et al. 2023), image captioning (Chen et al. 2022), object detection (Du et al. 2022), and segmentation (Yun et al. 2023)—benefiting applications in autonomous driving (Park et al. 2024) and medical imaging (Zhang et al. 2024b).

While VLMs bridge visual and textual domains, our work leverages their strength in extracting relational and contextual cues from single images. Trained on large-scale image-text data, VLMs accurately capture semantics and object layouts (Zhao et al. 2024). However, optimized for single-image captioning, they struggle to compare two images and detect subtle relational nuances, as noted in prior analyses (Lin et al. 2025).

Motivated by these limitations, our method leverages VLMs to extract relational context from individual images, supplementing visual-only change detectors. Instead of directly comparing paired images, we generate reasoning-aware text as auxiliary information. This fusion of high-level relational cues with low-level visual differences enables our approach to capture subtle change details and overall scene dynamics, leading to more accurate change captions.

Methodology

Figure 2 shows an overall architecture of CORTEX, a plug-and-play framework designed to enhance visual-only change captioning models by incorporating explicit compositional reasoning signals in textual form. CORTEX is composed of

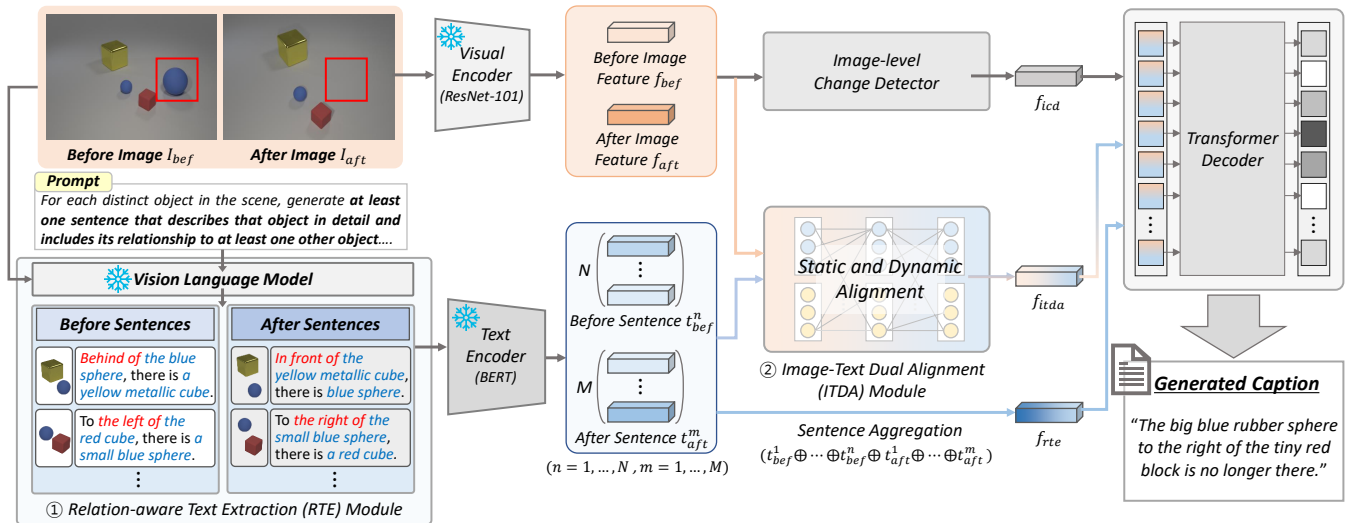


Figure 2: Overview of the proposed Compositional Reasoning-aware Text-guided (CORTEX) framework for change captioning, which combines the three modules. We introduce (1) Image-level change detector, which captures change cues between the two images; (2) RTE module, which extracts compositional reasoning sentence for each scene; and (3) ITDA module, which reinforces same-scene understanding for static alignment and identifies changes in dynamic alignment in cross-scene.

three modular components: (1) Image-level Change Detector, (2) Reasoning-aware Text Extraction (RTE) module, and (3) Image-Text Dual Alignment (ITDA) module. Given a pair of input images, “before” image I_{bef} and “after” image I_{aft} , a visual backbone (e.g., ResNet-101 (He et al. 2016)) extracts the corresponding visual features f_{bef} , f_{aft} . Then, the image-level change detector takes them as inputs and encodes f_{icd} , which captures low-level appearance differences between the two images. In parallel, the RTE module utilizes a VLM (e.g., InternVL2 (Chen et al. 2024)) to extract N compositional reasoning sentences T_{bef} from I_{bef} and M sentences T_{aft} from I_{aft} . Each sentence contains high-level compositional reasoning cues, such as relative attributes (e.g., comparisons of size or brightness) and spatial relations that are difficult to infer from pixel-level signal alone. A text encoder (e.g., BERT (Devlin et al. 2019)) embeds T_{bef} , T_{aft} into sentence features t_{bef} , t_{aft} . All the sentence features are concatenated to generate the RTE feature f_{rte} . The ITDA module aligns visual features with the textual cues at both within and across scenes, generating the text-augmented feature f_{itda} to support reasoning-aware scene understanding. The combination of compositional reasoning cues from RTE-generated texts and pixel-level visual features enables the model to effectively capture scene changes by complementing visual understanding with explicit textual reasoning. Finally, a transformer decoder generates change captions by integrating the outputs from all modules. More details are in the following subsections.

Reasoning-aware Text Extraction (RTE) Module

While existing image-level change detectors effectively capture appearance differences between two images, they often lack the ability to perform fine-grained contextual reasoning based on relative attributes and spatial context. To address

this limitation, we introduce the RTE module, which extracts structured, sentence-level descriptions specially designed to support compositional reasoning.

In the RTE module, we leverage a frozen VLM (e.g., InternVL2 (Chen et al. 2024)) to generate textual descriptions from I_{bef} and I_{aft} .

Rather than generating generic descriptions, we prompt the VLM to extract compositional reasoning cues that include semantic details, which are often missed by visual features alone. To extract high-quality compositional-reasoning sentences, we use the following prompt:

Prompt for Compositional Reasoning

*Analyze the image thoroughly. For each distinct object in the scene, generate **at least one sentence** that describes that object in detail and includes its relationship to at least one other object. Each sentence should mention **the object’s color, shape, size, and relevant spatial relationships** (such as distance, proximity, or grouping).*

This carefully crafted prompt encourages the generation of compositional reasoning cues by guiding the model to describe each object with detailed attributes (e.g., color, shape, size) and its spatial relationships (Qiu et al. 2025) with other objects. It also ensures consistency and completeness, which are crucial for structured analysis.

At this time, since each scene varies in object density and complexity, the number of extracted sentences is determined dynamically based on the scene, denoted as $T_{bef} = \{T_{bef}^n\}_{n=1}^N$, and $T_{aft} = \{T_{aft}^m\}_{m=1}^M$ (N sentences for ‘before’ image and M sentences for ‘after’ image).

Subsequently, the generated sentences T_{bef} and T_{aft} are

passed through a text encoder (Devlin et al. 2019) to produce the sentence features $t_{bef}, t_{aft} \in \mathbb{R}^c$ (c denotes the channel number). We will release all the generated sentences to support advanced research in change captioning. Detailed descriptions are provided in the supplementary materials.

Image-Text Dual Alignment (ITDA) Module

Although the RTE module extracts compositional-reasoning text from individual images, these features are embedded in a different latent space than visual features from the image-level change detector. To unify image and text modalities and fully exploit their complementary strengths, we introduce the ITDA module. The ITDA module consists of two components: (i) **static alignment**, which enhances compositional understanding within each scene, and (ii) **dynamic alignment**, which emphasizes changes across scenes.

Static Alignment. First, we design the static alignment to capture and refine intra-scene compositional structure. To do this, we align visual features with compositional-reasoning sentence features from the same scene (either “before” or “after”). As shown in Figure 3 (a), this process takes a visual feature f_{bef} or f_{aft} and the corresponding compositional-reasoning sentence feature t_{bef} or t_{aft} . For the “before” scene, t_{bef} provides compositional reasoning cues that reflect relative attributes and spatial relationships. We apply cross-attention between visual feature f_{bef} and each of the N compositional-reasoning sentence features, then average the outputs, which can be formulated as:

$$f_{bef}^{s(t \rightarrow i)} = \frac{1}{N} \sum_{n=1}^N \text{Attn}(t_{bef}^n, f_{bef}, f_{bef}), \quad (1)$$

where $\text{Attn}(Q, K, V) = \text{Softmax}(QK^\top / \sqrt{c})V$. This yields a text-augmented static feature for the “before” image. The same process is applied to the “after” scene, where its M compositional-reasoning sentence features are used to produce the text-augmented static feature $f_{aft}^{s(t \rightarrow i)}$.

To ensure semantic consistency between $f_{bef}^{s(t \rightarrow i)}$, $f_{aft}^{s(t \rightarrow i)}$ and the image-level change detector feature f_{icd} , we compute self-attended visual features as follows:

$$f_{bef}^{s(i \rightarrow i)} = \text{Attn}(f_{bef}, f_{bef}, f_{bef}), \quad (2)$$

$$f_{aft}^{s(i \rightarrow i)} = \text{Attn}(f_{aft}, f_{aft}, f_{aft}). \quad (3)$$

Then, we introduce a static alignment loss \mathcal{L}_{sa} to encourage both latent spaces to share common semantics, denoted as:

$$\mathcal{L}_{sa} = \frac{1}{2} (\|f_{bef}^{s(t \rightarrow i)} - f_{bef}^{s(i \rightarrow i)}\|_2^2 + \|f_{aft}^{s(t \rightarrow i)} - f_{aft}^{s(i \rightarrow i)}\|_2^2). \quad (4)$$

This loss guides the model to integrate compositional details into visual representations, thereby improving scene-level compositional reasoning through static alignment.

Dynamic Alignment. Further, to capture the cross-scenes changes, we devise the dynamic alignment by applying cross-attention between the visual features of one scene and

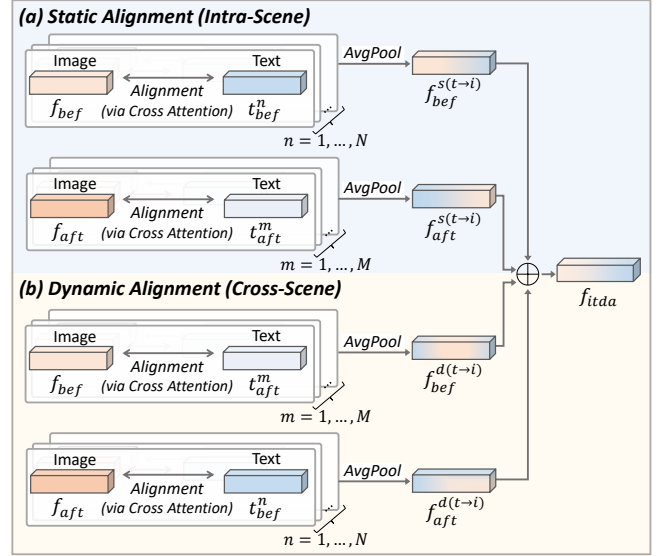


Figure 3: Overview of the ITDA module. (a) Static alignment matches each image with its corresponding compositional texts extracted by the RTE module. (b) Dynamic alignment matches each image with texts from the cross-scene to highlight the changes. \oplus denotes concatenation.

the compositional-reasoning sentence features of the other. As shown in Figure 3 (b), for the “before” scene, we use its visual feature f_{bef} with M “after” sentence feature t_{aft} . Then the outputs are averaged to produce the final text-augmented dynamic feature for the “before” scene:

$$f_{bef}^{d(t \rightarrow i)} = \frac{1}{M} \sum_{m=1}^M \text{Attn}(t_{aft}^m, f_{bef}, f_{bef}), \quad (5)$$

and “after” text-augmented dynamic feature $f_{aft}^{d(t \rightarrow i)}$ is computed through a similar process. To ensure that these text-augmented dynamic features align with the visual features, we design the dynamic alignment loss \mathcal{L}_{da} as:

$$f_{bef}^{d(i \rightarrow i)} = \text{Attn}(f_{aft}, f_{bef}, f_{bef}), \quad (6)$$

$$f_{aft}^{d(i \rightarrow i)} = \text{Attn}(f_{bef}, f_{aft}, f_{aft}), \quad (7)$$

$$\mathcal{L}_{da} = \frac{1}{2} (\|f_{bef}^{d(t \rightarrow i)} - f_{bef}^{d(i \rightarrow i)}\|_2^2 + \|f_{aft}^{d(t \rightarrow i)} - f_{aft}^{d(i \rightarrow i)}\|_2^2), \quad (8)$$

where $f_{bef}^{d(i \rightarrow i)}$ denotes cross-attended visual features.

The \mathcal{L}_{da} enforces alignment between the visual features from one scene and the text feature of another scene, facilitating more effective identification of scene differences by leveraging complementary multimodal information.

Finally, $f_{bef}^{s(t \rightarrow i)}$, $f_{aft}^{s(t \rightarrow i)}$, $f_{bef}^{d(t \rightarrow i)}$ and $f_{aft}^{d(t \rightarrow i)}$ are concatenated to generate the final feature f_{itda} . We devise alignment loss \mathcal{L}_{align} which combines \mathcal{L}_{sa} and \mathcal{L}_{da} , denoted as:

$$\mathcal{L}_{align} = \mathcal{L}_{sa} + \mathcal{L}_{da}. \quad (9)$$

As a result, ITDA module enables a unified understanding of both intra-scene composition and cross-scene differences

Method	Total Performance					Semantic Change				
	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
DUDA (ICCV'19)	47.3	33.9	-	112.3	24.5	42.9	29.7	-	94.6	19.9
DUDA+ (CVPR'21)	51.2	37.7	70.5	115.4	31.1	49.9	34.3	65.4	101.3	27.9
MCCFormers-D (ICCV'21)	52.4	38.3	-	121.6	26.8	-	-	-	-	-
MCCFormers-S (ICCV'21)	57.4	41.2	-	125.5	32.4	-	-	-	-	-
PCL w/o Pre-training (AAAI'22)	32.7	27.7	57.2	89.8	-	-	-	-	-	-
NCT (TMM'23)	55.1	40.2	73.8	124.1	32.9	53.1	36.5	70.7	118.4	30.9
I3N-TD (TMM'23)	55.8	40.6	73.9	125.6	32.8	-	-	-	-	-
VARD-Trans (TIP'23)	55.4	40.1	73.8	126.4	32.6	53.6	36.7	71.0	119.1	30.5
RDD+ACR (AAAI'25)	56.1	41.3	75.0	128.1	33.5	-	-	-	-	-
DECIDER (AAAI'25)	56.4	39.7	75.3	131.3	-	-	-	-	-	-
SCORER (ICCV'23)	56.3	41.2	74.5	126.8	33.3	54.4	37.6	71.7	122.4	31.6
CORTEX (SCORER)	57.0	42.7	75.9	128.8	33.9	54.9	39.2	74.0	127.5	32.8
SMART (TPAMI'24)	56.1	40.8	74.2	127.0	33.4	54.3	37.4	71.8	123.6	32.0
CORTEX (SMART)	56.5	42.1	75.7	130.2	34.0	54.6	39.1	74.5	130.3	32.9
DIRL (ECCV'24)	-	-	-	-	-	54.6	38.1	71.9	123.6	31.8
DIRL [†] (ECCV'24)	55.5	40.8	73.4	125.3	33.4	55.4	38.4	72.1	123.2	32.7
CORTEX (DIRL)	57.4	43.0	76.2	130.7	34.2	55.4	39.6	74.6	131.1	33.5

Table 1: Performance comparisons on the CLEVR-Change dataset (BLEU-4 (\mathcal{B}), METEOR (\mathcal{M}), ROUGE-L (\mathcal{R}), CIDEr (\mathcal{C}), and SPICE (\mathcal{S})). CORTEX consistently improves the performances of existing methods that provide publicly available source code. [†] denotes reproduced results of the method, as DIRL does not report ‘total’ performance.

by leveraging compositional-reasoning sentences. This dual alignment compensates for the limitation of visual-only approaches by providing semantically enriched and contextually grounded representations.

Building on this alignment, we define the total training objective of CORTEX as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cap} + \lambda \mathcal{L}_{align}, \quad (10)$$

where λ is balancing hyper-parameter, \mathcal{L}_{cap} represents the captioning loss that guides the overall caption generation process of the transformer decoder (Tu et al. 2024a).

Experiments

Datasets and Evaluation Metrics

Datasets. In our experiments, we use three datasets: (i) CLEVR-Change, (ii) CLEVR-DC, and (iii) Spot-the-Diff. **CLEVR-Change** (Park, Darrell, and Rohrbach 2019) is a large-scale synthetic dataset for controlled settings, consisting of 79,606 image pairs and 493,735 captions. It is divided into 67,660 training, 3,976 validation, and 7,970 test pairs.

CLEVR-DC (Kim et al. 2021) contains 48,000 image pairs with additional dynamic viewpoint shifts as distractors. We follow the official split: 85% training, 5% validation, and 10% testing.

Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) consists of 13,192 well-aligned surveillance image pairs, divided into 8:1:1 splits for training, validation, and testing, providing a real-world benchmark for generalization.

Evaluation Metrics. We evaluate caption quality using standard metrics: BLEU-4 (\mathcal{B}) (Papineni et al. 2002), METEOR (\mathcal{M}) (Banerjee and Lavie 2005), ROUGE-L (\mathcal{R}) (Lin 2004), CIDEr (\mathcal{C}) (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (\mathcal{S}) (Anderson et al. 2016).

Method	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
DUDA (ICCV'19)	40.3	27.1	-	56.7	16.1
M-VAM (ECCV'20)	40.9	27.1	-	60.1	15.8
VACC (ICCV'21)	45.0	29.3	-	71.7	17.6
NCT (TMM'23)	47.5	32.5	65.1	76.9	15.6
VARD-Trans (TIP'23)	48.3	32.4	-	77.6	15.4
SCORER (ICCV'23)	49.4	33.4	66.1	83.7	16.2
CORTEX (SCORER)	52.2	34.0	67.3	88.7	16.5
SMART [†] (TPAMI'24)	47.9	32.7	65.4	82.9	15.6
CORTEX (SMART)	53.2	32.9	66.6	86.6	16.9
DIRL (ECCV'24)	51.4	32.3	66.3	84.1	16.8
CORTEX (DIRL)	55.3	32.9	67.8	89.7	17.0

Table 2: Performance comparisons with state-of-the-art methods on the CLEVR-DC dataset.

Implementation Details

We adopt three state-of-the-art baselines with publicly available code: SCORER, SMART, and DIRL (Tu et al. 2023b, 2024b,a) as our image-level change detectors. At this stage, following the standard protocol in change captioning, we use a pre-trained ResNet-101 (He et al. 2016) to extract features from image pairs. We use InternVL2-8B (Chen et al. 2024) as the VLM to generate compositional reasoning texts.

We set the number of attention heads to $h = 8$ and $\lambda = 10^{-3}$ for SCORER and $\lambda = 10^{-4}$ for both SMART and DIRL in Eq. (10). CORTEX is trained using Adam optimizer (Kingma and Ba 2017) on a single RTX 4090 GPU.

Comparison with Existing Methods

Results on the CLEVR-Change Dataset. We compared our method with SOTA methods on the CLEVR-Change

Method	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
DUDA+ (CVPR’21)	8.1	12.5	-	34.5	-
MCCFormers-D (ICCV’21)	10.0	12.4	-	43.1	18.3
MCCFormers-S (ICCV’21)	-	12.3	-	41.6	16.3
I3N-TD (TMM’23)	-	13.0	31.5	42.7	18.6
VARD-Trans (TIP’23)	-	12.5	29.3	30.3	17.3
RDD+ACR (AAAI’25)	9.2	13.9	31.0	43.6	-
DECIDER (AAAI’25)	10.7	14.2	41.6	39.9	-
SCORER (ICCV’23)	10.2	12.2	-	38.9	18.4
CORTEX (SCORER)	10.5	12.6	33.2	40.3	19.4
SMART (TPAMI’24)	-	13.5	31.6	39.4	19.0
CORTEX (SMART)	9.5	12.2	32.7	41.0	19.0
DIRL (ECCV’24)	10.3	13.8	32.8	40.9	19.9
CORTEX (DIRL)	11.6	13.9	33.4	49.5	21.4

Table 3: Performance comparisons with state-of-the-art methods on the Spot-the-Diff dataset.

RTE	ITDA	Total				
		\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
Baseline (DIRL)	-	55.5	40.8	73.4	125.3	33.4
✓	-	55.8	41.6	74.8	128.5	33.9
✓	✓	57.4	43.0	76.2	130.7	34.2

Table 4: Effect of the proposed modules (RTE and ITDA) on the CLEVR-Change dataset.

dataset. We evaluated performance for both ‘total’ and ‘semantic change’ settings (Tu et al. 2024a), where ‘semantic change’ refers to cases with actual changes, and ‘total’ includes both changed and unchanged cases.

As shown in Table 1, incorporating compositional reasoning through CORTEX into three SOTAs (SCORER, SMART, DIRL) consistently improves performance. This shows that CORTEX effectively complements visual-only methods by injecting explicit compositional reasoning into the change captioning process. The consistent gains across baselines confirm its effectiveness and the importance of compositional understanding.

Results on the CLEVR-DC Dataset. We evaluated performance on the CLEVR-DC dataset to assess robustness under extreme viewpoint changes. As shown in Table 2, integrating CORTEX into three SOTA methods consistently achieved the best results across all metrics. By incorporating VLM-extracted textual cues, CORTEX enhances image-text understanding and enables robust compositional reasoning even under drastic viewpoint shifts.

Results on the Spot-the-Diff Dataset. Further, to validate the generalization ability of our method in real-world scenes, we conducted experiments on the Spot-the-Diff dataset, which consists of image pairs from surveillance cameras. In Table 3, applying CORTEX yields superior performance across most metrics. These results emphasize the robustness of our architecture, showing its ability to generalize across diverse real-world scenes and change contexts.

\mathcal{L}_{align}		Total				
\mathcal{L}_{sa}	\mathcal{L}_{da}	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
✗	✗	56.6	41.5	75.1	127.9	33.5
✓	✗	56.3	41.8	75.5	128.4	34.0
✗	✓	56.6	41.8	75.6	128.9	33.7
✓	✓	57.4	43.0	76.2	130.7	34.2

Table 5: Effect of the proposed static loss (\mathcal{L}_{sa}) and dynamic loss (\mathcal{L}_{da}) on the CLEVR-Change dataset.

Prompt Type	\mathcal{B}	\mathcal{M}	\mathcal{R}	\mathcal{C}	\mathcal{S}
Generic descriptions	56.5	41.6	75.3	129.5	33.5
Compositional reasoning	57.4	43.0	76.2	130.7	34.2

Table 6: Effect of different VLM prompt types on the CLEVR-Change dataset.

Ablation Studies

We conducted ablation studies to analyze the effect of (i) the proposed modules, (ii) the proposed losses, and (iii) different VLM prompt types. All variants were evaluated under the ‘total’ setting on the CLEVR-Change dataset, using DIRL (Tu et al. 2024a) as the baseline.

Effect of the Proposed Modules. We evaluated the impact of the two proposed modules, RTE and ITDA. Table 4 shows that applying the RTE module, which leverages compositional reasoning text significantly outperforms the visual-only baseline (DIRL). The best performance is achieved by using both modules. Note that ITDA was not tested alone, as it requires the compositional reasoning text generated by the RTE module.

Effect of the Proposed Losses. Table 5 shows the effect of the two losses, \mathcal{L}_{sa} and \mathcal{L}_{da} , used in \mathcal{L}_{align} of the ITDA module. Adding either loss individually led to improvements over the baseline without both losses. We achieved the highest performance when both losses were considered.

Effect of Prompt Types for VLM. To investigate the impact of reasoning cues on the visual-only method, we compare two types of VLM-generated text: (1) generic descriptions prompted with “Analyze the image and list sentences that describe the scene.” and (2) compositional reasoning sentences guided to include relative attributes and inter-object relationships. In Table 6, compositional reasoning consistently yields better performance, highlighting the importance of structured cues for inferring fine-grained relational and attribute information essential to describing changes.

Qualitative Results

We compared CORTEX with DIRL in Figure 4. CORTEX better captures compositional structures, generating outputs more closely aligned with the ground truth (see [blue / red](#) parts). This demonstrates the strength of compositional reasoning, where textual guidance enables the model to produce

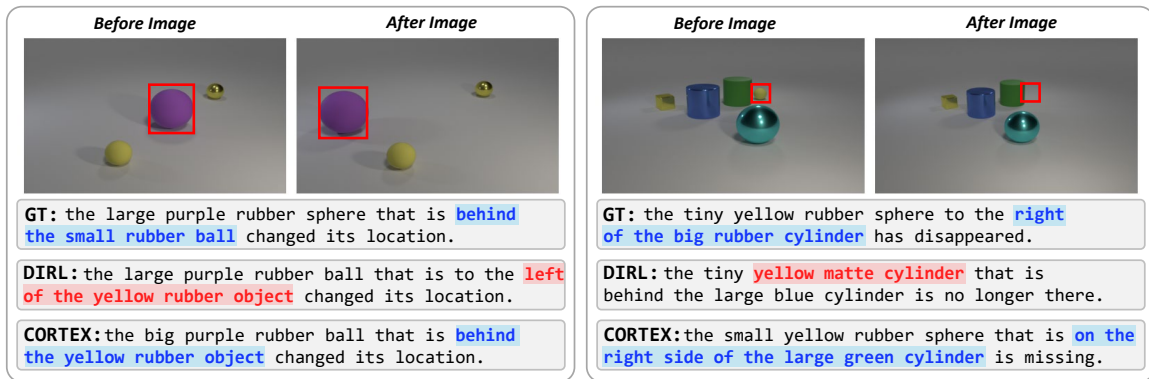


Figure 4: Visualization examples in the CLEVR-Change dataset (Blue / red : correct/incorrect compositional reasoning cues).

VLM	B	M	R	C	S
Baseline	55.5	40.8	73.4	125.3	33.4
LLaVA (Liu et al. 2024)	56.9	42.3	75.7	130.0	34.1
InternVL2 (Chen et al. 2024)	57.4	43.0	76.2	130.7	34.2

Table 7: Comparison of different VLMs used in our method on the CLEVR-Change dataset. The top row shows the visual-only baseline (DIRL) without text-based guidance.

VLM Usage	B	M	R	C	S
Direct Prediction (VLM for Paired Images)	2.7	10.7	21.0	12.3	12.5
Auxiliary Context (Ours) (VLM for Single Image)	11.6	13.9	33.4	49.5	21.4

Table 8: Comparison of VLM usage strategies on the Spot-the-Diff dataset. Our method enhances visual-only change captioning with compositional reasoning from single-image captions, outperforming direct paired-image approach.

more precise and context-aware change descriptions.

Discussion

Effect of VLM Variants on CORTEX. While InternVL2 is our primary VLM, we also evaluated CORTEX with LLaVA (Liu et al. 2024). As shown in Table 7, CORTEX consistently outperforms the baseline, effectively leveraging image-text understanding across different VLMs.

Comparison of VLM Usage Strategies. As VLMs are trained to understand semantics and spatial layouts from *single images*, they often *struggle to compare two images and detect subtle differences*, as noted in (Lin et al. 2025).

To investigate this issue, we compared two VLM usage strategies on the Spot-the-Diff dataset (Table 8). The first approach directly infers changes by feeding paired “before” and “after” images into the VLM with a prompt to identify changes (in the supplementary). In contrast, second approach (Ours) leverages the single-image reasoning ability of VLMs by extracting compositional reasoning sentences

Method	Offline	Online		
	VLM (per img)	Train. (per iter)	Infer. (per img)	#Learnable params
DIRL (ECCV’24)	-	0.77s	0.007s	14.4M
CORTEX (DIRL)	3.94s	0.79s	0.008s	18.2M

Table 9: Offline/online time analysis with baseline DIRL.

from individual images and incorporating them as auxiliary cues into visual-only methods (Tu et al. 2024a). These sentences provide fine-grained semantic cues (*e.g.*, object attributes and spatial relations) often missed by visual-only models, enabling CORTEX to better capture subtle changes and compositional context.

Computational Costs. Table 9 compares offline captioning time, online training/inference times, and learnable parameters. Following (Wei et al. 2024), captions are generated with the VLM (InternVL2) offline before training, reducing runtime overhead. While VLM-based captioning adds one-time preprocessing step, its impact on efficiency is minimal, keeping our method lightweight and compositional-aware.

Limitations. While our framework achieves strong performance by utilizing compositional reasoning cues from VLM-generated text, the use of VLMs incurs computational overhead. For more practical applications, our future work will aim to reduce this overhead.

Conclusion

We introduced CORTEX, a new framework for change captioning that incorporates textual compositional reasoning cues while preserving the strengths of image-level change detectors. To this end, we devise two modules: RTE for extracting compositional cues and ITDA for aligning image-text features to capture both static and dynamic changes. Designed as a plug-and-play component, CORTEX can be seamlessly integrated into existing methods. Experimental results highlight the importance of explicit compositional reasoning in accurately describing scene changes.

Acknowledgments

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by IITP grant funded by the Korea government (MSIT) (No. RS-2025-25442384, IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program, IITP-2022-II220078: Explainable Logical Reasoning for Medical Knowledge Generation) and conducted by CARAI grant funded by DAPA and ADD (UD230017TD).

References

- Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, 382–398. Springer.
- Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.
- Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.
- Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hoxha, G.; Chouaf, S.; Melgani, F.; and Smara, Y. 2022. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.
- Hwang, I.; Kim, H.; and Kim, Y. M. 2023. Text2scene: Text-driven indoor scene stylization with part-aware details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1890–1899.
- Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.
- Khan, Z.; BG, V. K.; Schuler, S.; Yu, X.; Fu, Y.; and Chandraker, M. 2023. Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images! In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15005–15015.
- Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Viewpoint-Agnostic Change Captioning with Cycle Consistency. In *ICCV*.
- Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.
- Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.
- Li, R.; Li, L.; Zhang, J.; Zhao, Q.; Wang, H.; and Yan, C. 2025. Region-aware Difference Distilling with Attribute-guided Contrastive Regularization for Change Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4887–4895.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- Lin, Q.; He, K.; Zhu, Y.; Xu, F.; Cambria, E.; and Feng, M. 2025. Cross-Modal Knowledge Diffusion-Based Generation for Difference-Aware Medical VQA. *IEEE Transactions on Image Processing*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4624–4633.
- Park, S.; Lee, M.; Kang, J.; Choi, H.; Park, Y.; Cho, J.; Lee, A.; and Kim, D. 2024. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 980–987.
- Qiu, H.; Gao, M.; Qian, L.; Pan, K.; Yu, Q.; Li, J.; Wang, W.; Tang, S.; Zhuang, Y.; and Chua, T.-S. 2025. STEP: Enhancing Video-LLMs’ Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3284–3294.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Tu, Y.; Li, L.; Su, L.; Lu, K.; and Huang, Q. 2023a. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25: 9518–9529.

- Tu, Y.; Li, L.; Su, L.; Yan, C.; and Huang, Q. 2024a. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *European Conference on Computer Vision*, 311–328. Springer.
- Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024b. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; Yan, C.; and Huang, Q. 2023b. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2805–2815.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
- Wei, L.; Lang, C.; Chen, Z.; Wang, T.; Li, Y.; and Liu, J. 2024. Generated and pseudo content guided prototype refinement for few-shot point cloud segmentation. *Advances in Neural Information Processing Systems*, 37: 31103–31123.
- Yun, S.; Park, S. H.; Seo, P. H.; and Shin, J. 2023. If-seg: Image-free semantic segmentation via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2967–2977.
- Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5625–5644.
- Zhang, J.; Wang, G.; Kalra, M. K.; and Yan, P. 2024b. Disease-informed Adaptation of Vision-Language Models. *IEEE Transactions on Medical Imaging*, 1–1.
- Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; and Chang, B. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *ICLR*.
- Zhong, G.; Hu, J.; Chen, J.; Yuan, J.; and Pan, W. 2025. DECIDER: Difference-aware Contrastive Diffusion Model with Adversarial Perturbations for Image Change Captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10662–10670.