ViewDelta: Scaling Scene Change Detection through Text-Conditioning

Supplementary Material

Subin Varghese
University of Houston
4226 MLK Blvd Houston TX
srvargh2@cougarnet.uh.edu

Joshua Gao
University of Houston
4226 MLK Blvd Houston TX
jkgao@cougarnet.uh.edu

Vedhus Hoskere
University of Houston
4226 MLK Blvd Houston TX
vhoskere@central.uh.edu

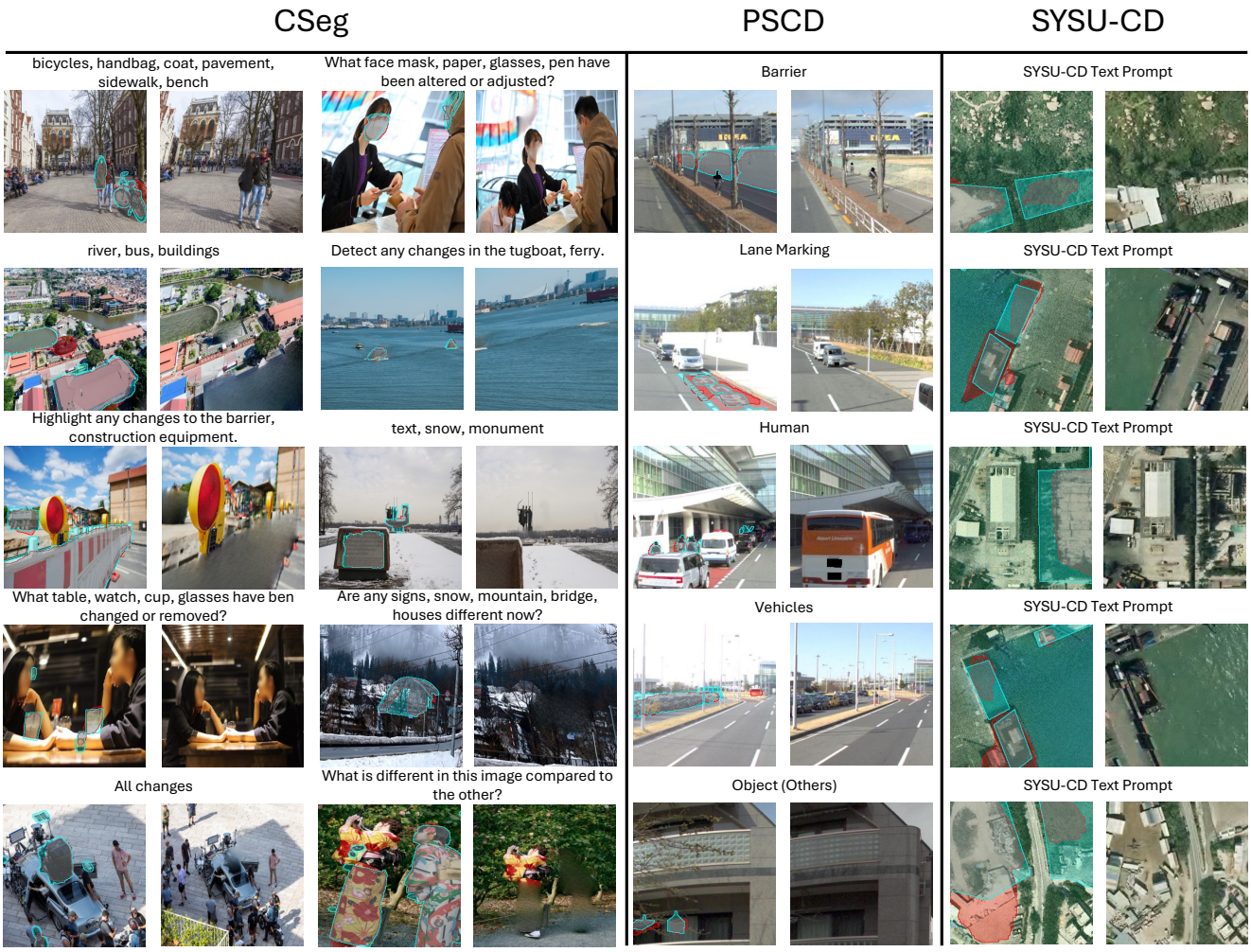

Figure 1. **Qualitative Results**: We show predictions from the jointly trained ViewDelta model in red, along with the ground truth in blue. **The SYSU-CD Text Prompt is** "urban development, suburban expansion, pre-construction groundwork, vegetation alteration, road widening, and coastal construction".

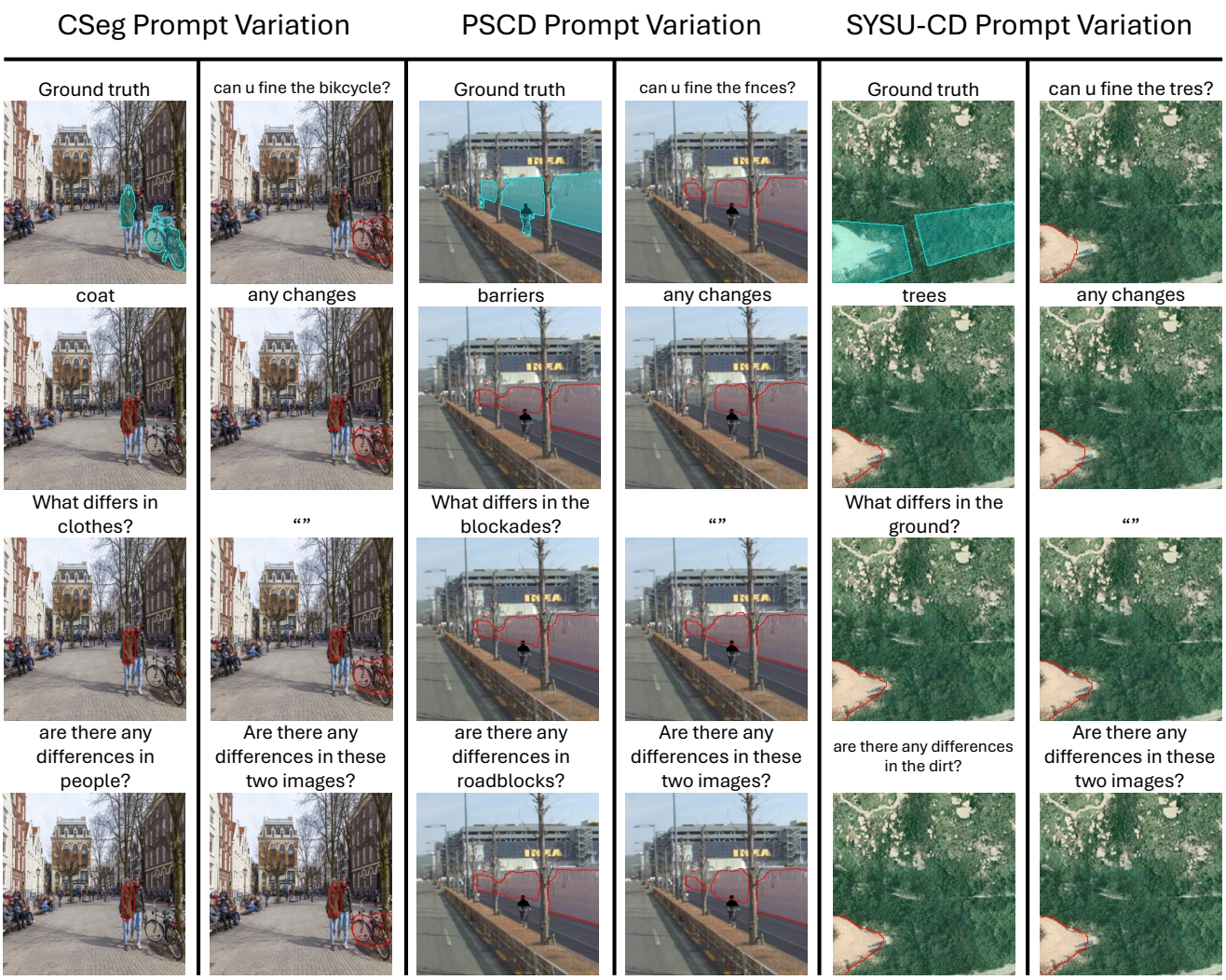

Figure 2. Effect of prompt variations on ViewDelta's performance across CSeg, PSCD, and SYSU-CD datasets. All images shown are Image A; corresponding Image B pairs are presented in Figure 1. The model demonstrates robustness to semantic equivalents, misspellings, and interrogative formulations on CSeg and PSCD, but shows limited generalization to novel prompts on SYSU-CD satellite imagery.

1. Effect of Prompt Variation

Figure 1 shows ViewDelta's predictions on image pairs across CSeg, PSCD, and SYSU-CD datasets, demonstrating the model's change detection capabilities. To further analyze the model's linguistic flexibility, Figure 2 presents the before-change images used to evaluate ViewDelta's robustness to prompt variations. We systematically test semantic equivalents, misspellings, and interrogative variations that CSeg incorporates by design to qualitatively evaluate the variations in change detection.

1.1. CSeg Dataset Analysis

On CSeg images, ViewDelta demonstrates remarkable robustness to prompt variations:

Semantic Equivalents: Despite training with the prompt "coat," the model successfully segments the same regions when prompted with semantically related terms like "clothes" and "people." This suggests that the frozen SigLip text encoder effectively captures semantic relationships between related concepts, enabling generalization beyond the exact training vocabulary.

Misspellings and Informal Language: The model maintains performance even with significant spelling errors and in-

formal language, as evidenced by the prompt "can u fine the bikcycle?" (containing both misspellings and informal text). This robustness indicates that the text encoder's pre-training on diverse internet text provides resilience to common linguistic variations.

Comprehensive Change Detection: When prompted for "all" changes using various phrasings ("All changes," "What is different?", "Find any differences"), ViewDelta consistently identifies the same set of changes, demonstrating understanding of the intent behind different formulations of comprehensive change requests.

1.2. PSCD Dataset Analysis

For PSCD street-view images, ViewDelta maintains consistent performance across prompt variations:

Class-Specific Prompts: The model accurately segments changes for specific semantic classes (e.g., "barrier," "lane marking," "vehicle") regardless of prompt phrasing, successfully adapting from the dataset's original eight-class structure.

Limitations with "All" Changes: The model occasionally misses subtle changes when prompted for all changes, particularly with people/cyclists who appear visually similar between frames. This suggests that the model may prioritize more visually distinct changes when not given specific class guidance, potentially due to the dataset's training emphasis on specific semantic categories rather than comprehensive change detection.

1.3. SYSU-CD Dataset Analysis

ViewDelta shows limited adaptation to novel prompts on SYSU-CD satellite imagery:

Domain-Specific Constraints: When prompted with non-satellite terminology (e.g., street-level object classes), the model struggles to map these concepts to satellite imagery features. This limitation stems from the training approach, where SYSU-CD was consistently paired with a single, domain-specific prompt: "urban development, suburban expansion, preconstruction groundwork, vegetation alteration, road widening, and coastal construction."

Cross-Domain Prompt Transfer: The inability to transfer prompts across domains (street-view to satellite) highlights a key limitation in the current training strategy. Unlike CSeg and PSCD, which benefit from diverse prompt-image associations, SYSU-CD's fixed prompt prevents the model from learning flexible mappings between varied text descriptions and satellite imagery changes.

Dataset Composition Impact: The absence of satellite imagery in CSeg's synthetic generation pipeline means that the model lacks exposure to satellite-specific visual features paired with various prompts. This gap in training data prevents the establishment of robust cross-domain prompt understanding.

1.4. Key Findings

Our analysis reveals three critical insights on text-conditioned change detection.

1. **Prompt Robustness:** ViewDelta successfully handles linguistic variations (misspellings, synonyms, question forms) when the visual domain matches training data, demonstrating the effectiveness of frozen pre-trained text encoders.

2. **Semantic Generalization:** The model generalizes well to semantically related concepts within familiar visual domains, suggesting effective vision-language alignment for in-domain applications.

3. **Cross-Domain Limitations:** Fixed prompt-domain associations during training limit cross-domain prompt transfer, indicating the need for more diverse prompt-image pairings across all visual domains in future training strategies.

These findings suggest that while ViewDelta achieves robust text-conditioned change detection within familiar domains, expanding prompt diversity across all training domains would further enhance its generalization capabilities.