A. Additional Qualitative Results

In this section, we provide results for more datasets. Figs. 8, 9 are results for image outpainting and Figs. 10, 11 and 12 are results for wide-range image blending. Lastly, we include results for wide-range image blending for wider gaps in Fig. 13.

B. Comparison of image completion according to caption

In this section, we include qualitative comparisons of different image captioning models and text-guided image manipulation models. The Blended Diffusion Model is a text-driven blended diffusion model that uses a clip-guidance classifier. In denoising diffusion process, an image corresponding to the input prompt is added to the masked part, and as the denoising step progresses, the original image and the added image are naturally blended to create an output image. When the OFA caption used in the study is applied to this blended diffusion model, the results are unsatisfactory. In Fig. 14, the output images of the blended diffusion model using the OFA caption model demonstrate that the results of GLIDE-implemented method generate natural and realistic images. Finally, Fig. 15 compare outcome images using captions by OFA and ClipCap. For the text-guided image manipulation model, GLIDE was utilized. The results show how much the results of the text-guided image manipulation model are dependent on the captions.

OFA is an unified multitask framework for multimodal learning, which is composed of an encoder E and a decoder D, each of which uses the Transformer as the backbone. Through multimodal learning, OFA performs various multimodal tasks with a unified framework. During inference, E extracts image features and D performs language modeling in case of image captioning tasks.

GLIDE is a guided diffusion based model for text-guided image synthesis and manipulation tasks. It utilizes two guidance methods: classifier-free guidance and CLIP guidance, both of which support free-form text prompts. For the first guidance method, a separate classifier which has to be trained is required. In order to implement a classifier-free guidance method, during training, class labels for class-conditional diffusion model are replaced by an empty sequence. The second method utilizes CLIP which is a joint representations learning method between texts and images. It provides a score which measures the distance between an image and a caption. To implement the second method, the classifier model of a classifier-guidance method is replaced by a noise-aware CLIP model, which was trained on noise data.

C. Code Descriptions

Our code is based on the PyTorch version of OFA and GLIDE. OFA is utilized to generate text hints, and GLIDE, for text-guided image manipulation. For GLIDE we set guidance_scale = 5.0, and upsample_temp = 0.997.

Codes for our project are available at
`https://github.com/Jihyun0510/caption_guided_extensive_painting.git`

D. CIDEr

In the experiment, as for the evaluation metric for self-critical sequence training (SCST), we utilized Consensus-based Image Description Evaluation (CIDEr). CIDEr measures the similarity of the predicted sentence c for the masked image I_{IC} to a set of reference or the ground-truth sentences $S = \{s_1, ..., s_m\}$ for the ground-truth image I_{GT}. The CIDEr score for n-grams of length n accounts for the average cosine similarity between the candidate sentence and a set of reference sentences. n-grams is a set of words, which is used to represent a sentence. The CIDEr score is computed as follows:

$$CIDEr(c, S) = \frac{1}{m} \sum_{i} \frac{g^n(c) \cdot g^n(s_i)}{\|g^n(c)\| \, \|g^n(s_i)\|}, \qquad (1)$$

where $g^n(s_i)$ is the vector of $g_k(s_i)$ for n sized n-grams and $\|g^n(\cdot)\|$ is the magnitude of $g^n(\cdot)$. The Term Frequency Inverse Document Frequency (TF-IDF) weight $g_k(s_i)$ measures the number of occurrences of n-gram ω_k in the reference sentence s_i. Similarly, $g^n(c)$ is the vector of $g_k(c_i)$ for n sized n-grams, which measures the occurrence times for predicted sentence c.

$$g_k(s_i) = \frac{h_k(s_i)}{\sum_{\omega_l \in \Omega} h_l(s_i)} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right), \qquad (2)$$

where Ω is the vocabulary of all n-grams, and I, the set of all images in the dataset. We combine the scores from n-grams of varying lengths as follows:

$$CIDEr(c, S) = \sum_{n=1}^{N} \omega_n CIDEr_n(c, S), \qquad (3)$$

where ω_n denotes the uniform weight for CIDEr scores.

E. Evaluation Metrics

In order to evaluate the results, we used reference IQA, specifically metrics which compare images by features. Detailed descriptions for each metric are as follows.

E.1. Fréchet Inception Distance (FID)

FID compares the distribution of features for two image datasets. It measures the distance between the two distribu-

tions, which are assumed to follow a Gaussian distribution.

$$\text{FID} = \|m - m_w\|_2^2 + Tr\left(C + C_w - 2(CC_w)^{1/2}\right), \quad (4)$$

where m and C are the mean and co-variances of the inception embeddings for real-data, and m_w and C_w are the mean and covariance matrices of the inception embeddings for the generated samples. The FID correlates well with image quality, and is capable of detecting mode collapse.

E.2. Kernel Inception Distance (KID)

Similar to FID, KID compares the distribution of features of two image datasets. It measures the squared Maximum Mean Discrepancy (MMD) between inception hidden layer activations.

$$MMD(p,q) = E_{x,x' p}[K(x,x^{'})]$$

$$+E_{x,x' q}[K(x,x^{'})] - 2E_{xp,x' q}[K(x,x^{'})], \quad (5)$$

where k is the default polynomial kernel. The final equation for KID is as follows:

$$KID = MMD\,(p,q)^2\,, \quad (6)$$

where x and q are extracted features from real and fake images.

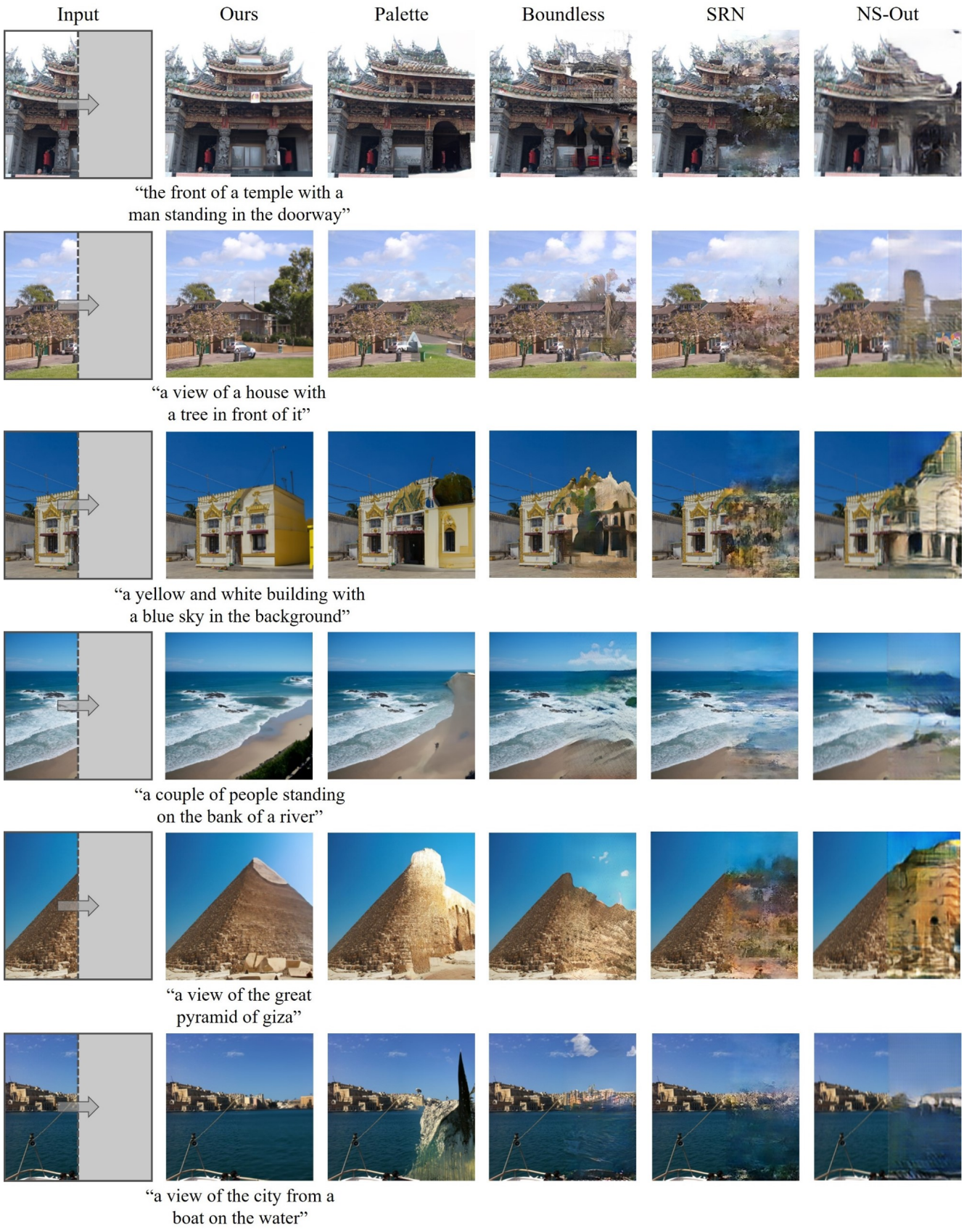

Figure 8. Image outpainting: Qualitative results for conventional and proposed methods on Landmarks dataset.

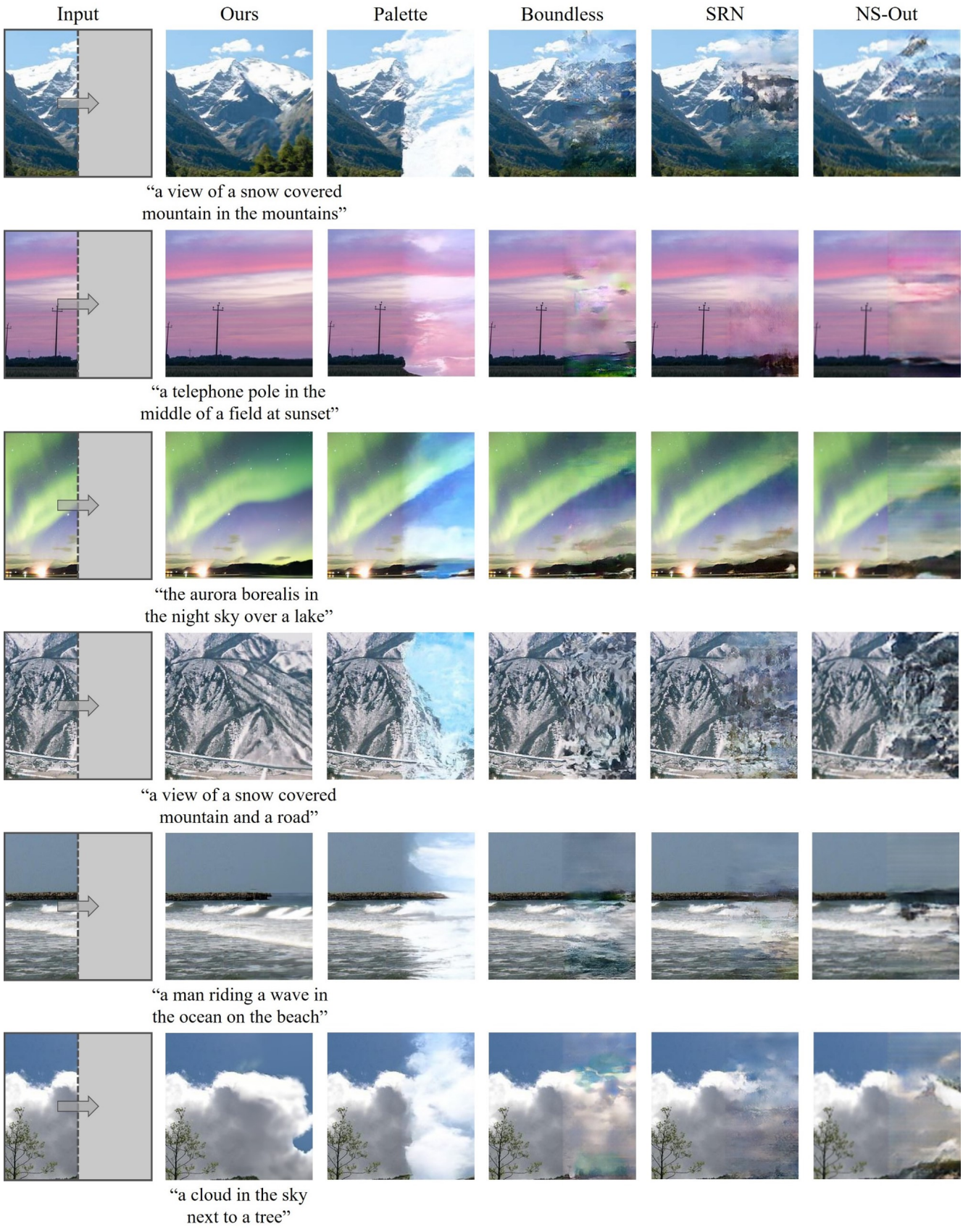

Figure 9. Image outpainting: Qualitative results for conventional and proposed methods on landscape dataset.

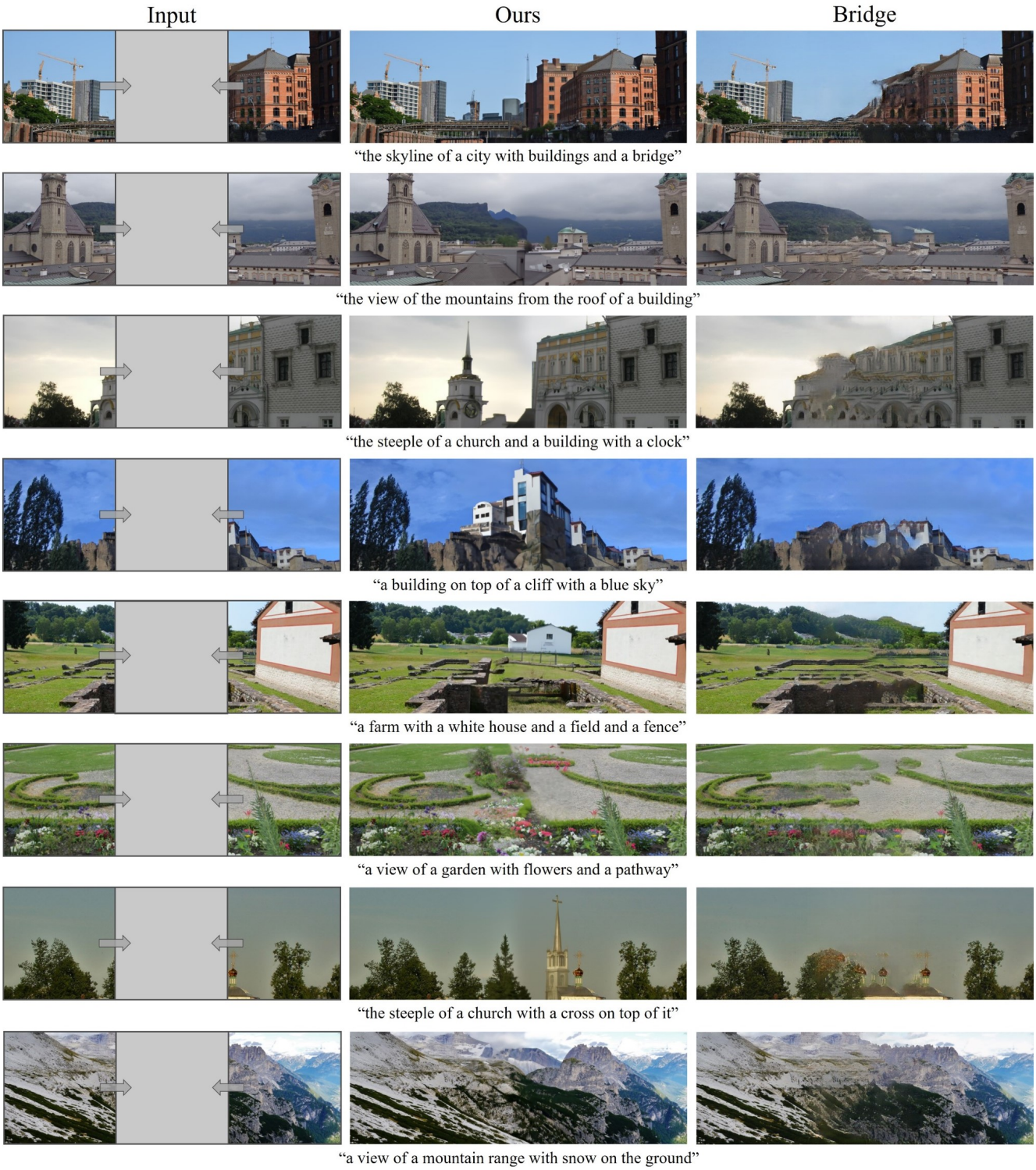

Figure 10. Wide-range image blending: Qualitative results for conventional and proposed method on Landmarks dataset.

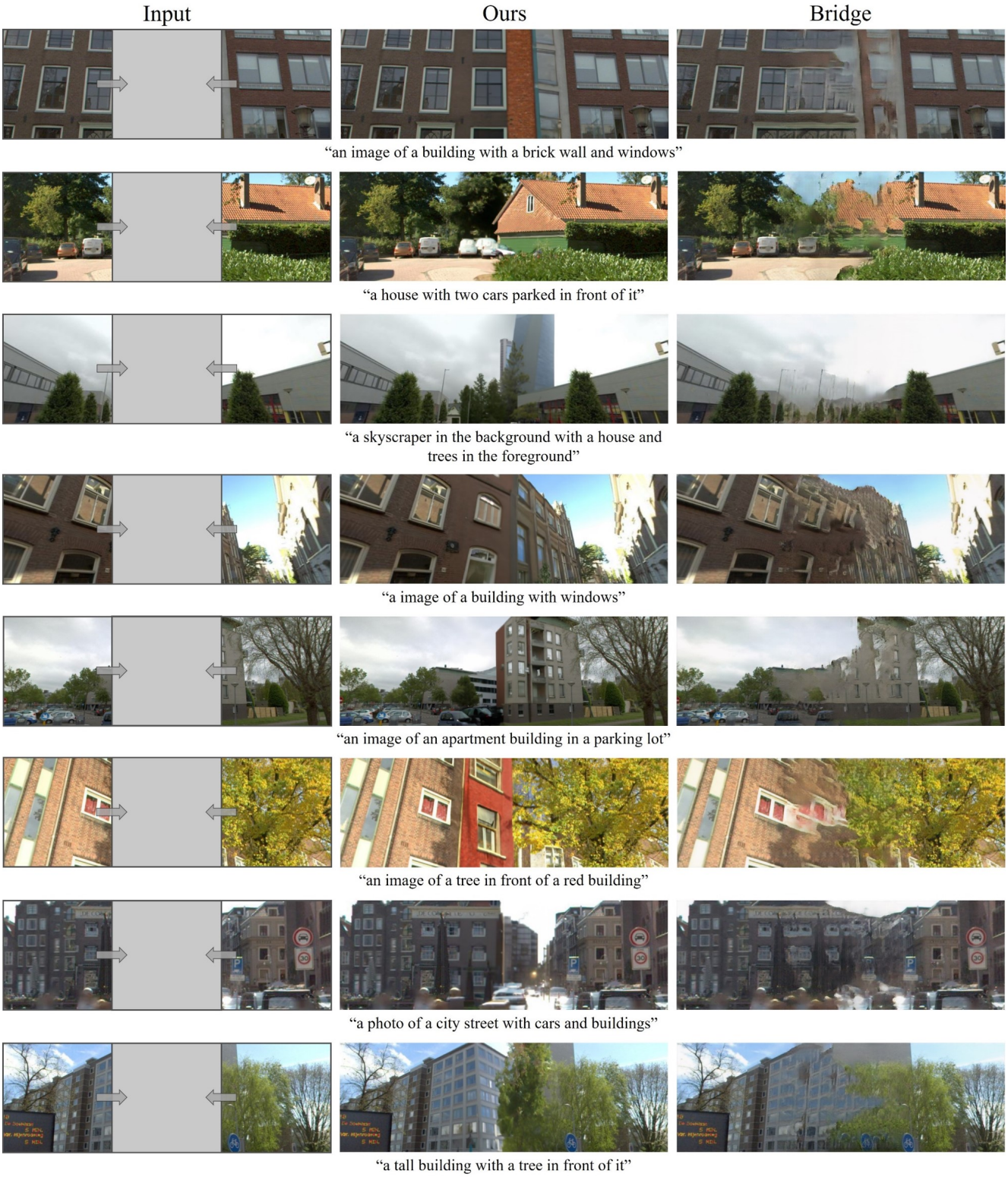

Figure 11. Wide-range image blending: Qualitative results for conventional and proposed method on AmsterTime dataset.

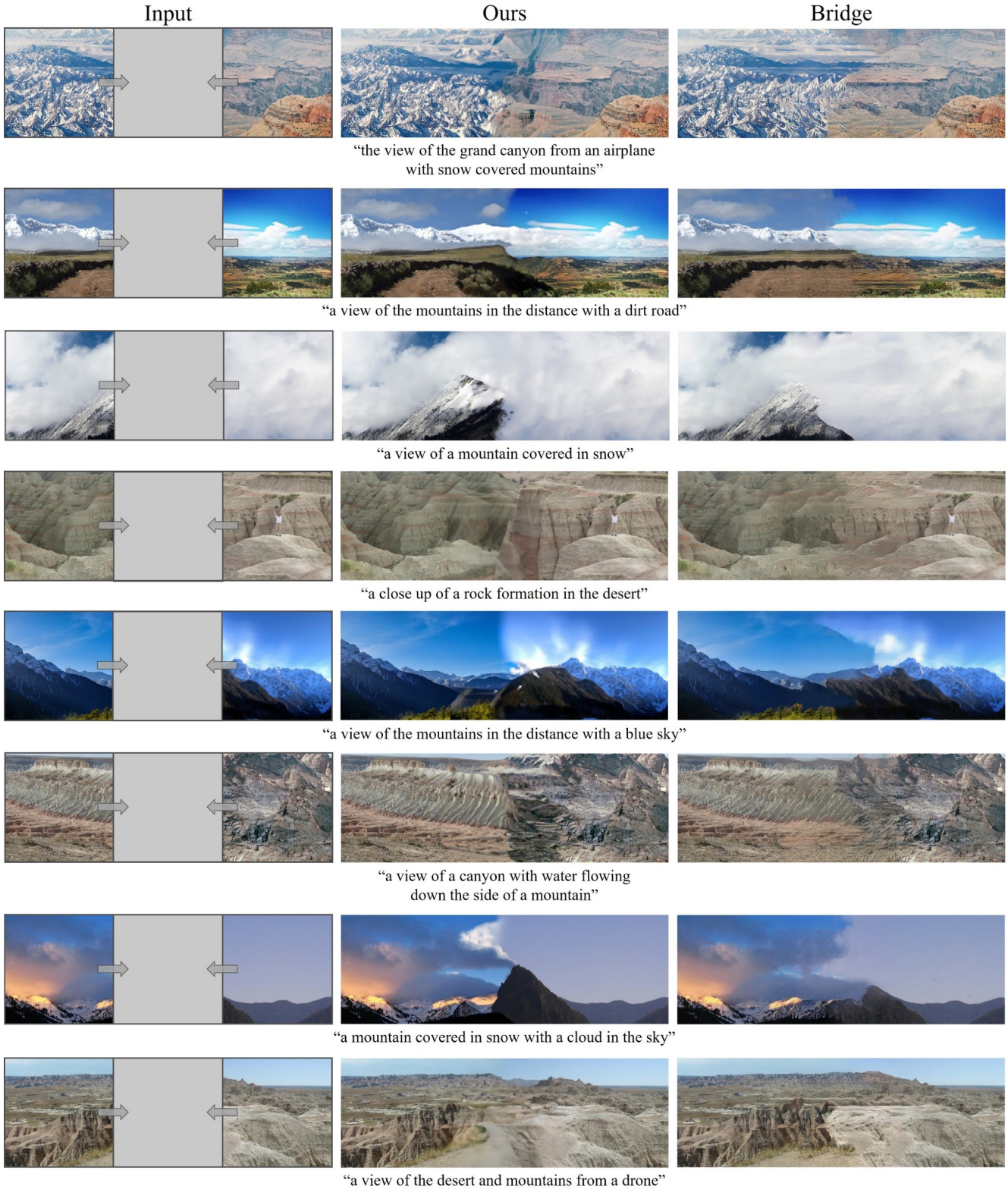

Figure 12. Wide-range image blending: Qualitative results for conventional and proposed method on Scenery dataset.

Figure 13. Panoramic image generated proposed CEP module. It is the result of applying the image blending task in the center after performing the outpainting task three times from the input images at both ends.

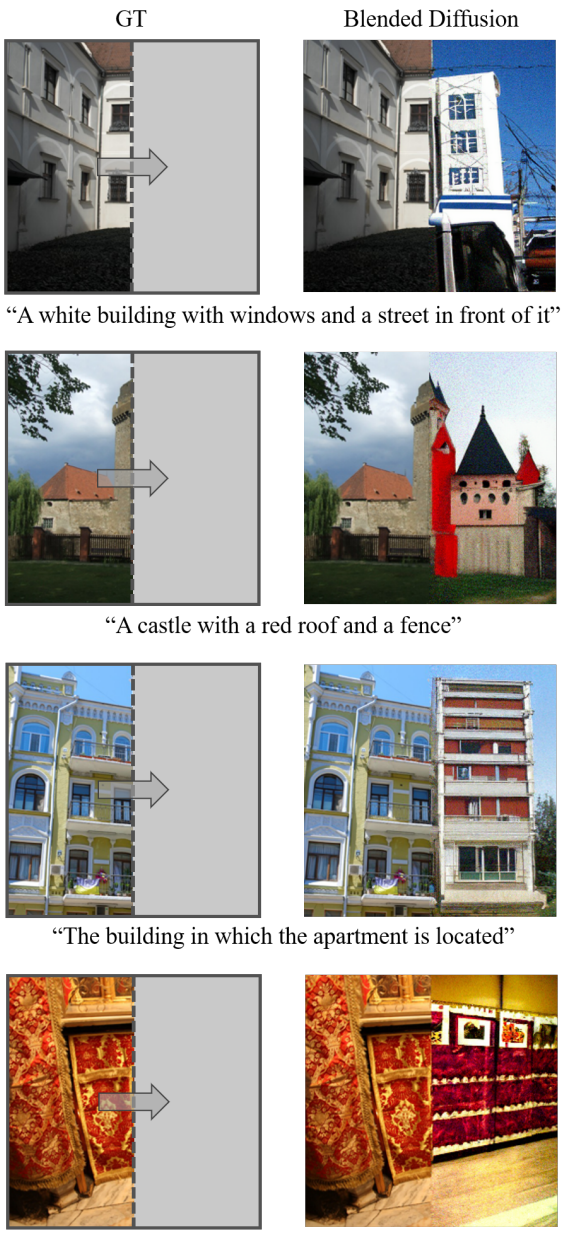

Figure 14. Image of Blended Diffusion model according to OFA caption.

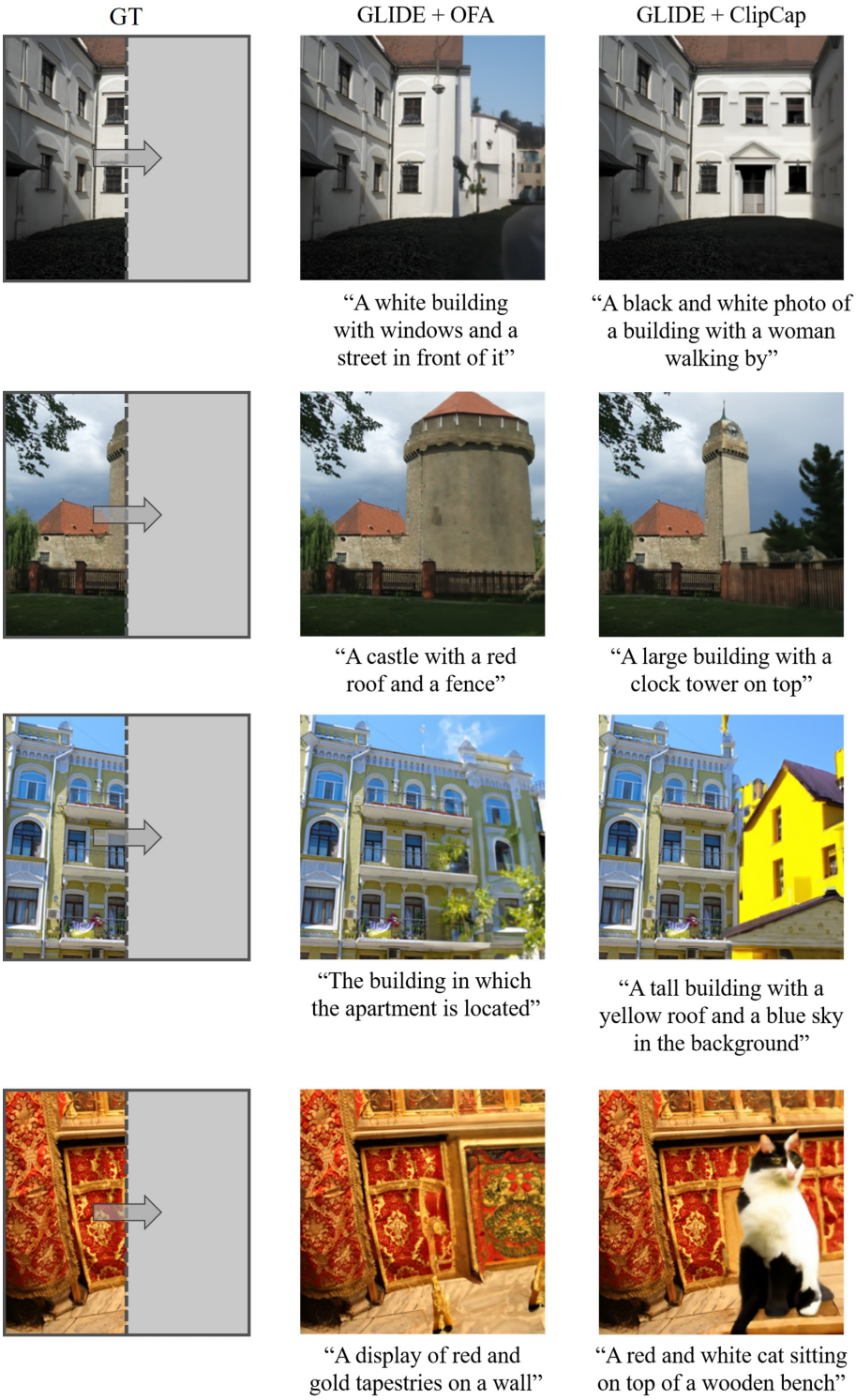

Figure 15. Comparison of OFA and ClipCap captionings with GLIDE