A. Detailed Algorithm

Algorithm 1 GLEAM Framework

Input: Image-text pair (v, t), target model M, perturbation bounds ϵ_v, ϵ_t, momentum coefficient μ, step size α, number of iterations T, numbers of transformations N, K

Output: adversarial example pair (v^{adv}, t^{adv})

// Stage 1: Visual adversarial example Generation
Initialize $v_0^{adv} \leftarrow v$, $g_0 \leftarrow 0$, $V^{adv} \leftarrow \{v_0^{adv}\}$
for $i = 0$ to $T - 1$ **do**
  Compute $\tilde{g}_i$ using Eq. 13
  Update g_{i+1} using Eq. 12
  Update v_{i+1}^{adv} using Eq. 11
  $V^{adv} \leftarrow V^{adv} \cup \{v_{i+1}^{adv}\}$
end for
// Stage 2: Text adversarial example Generation
for each word w_i in t **do**
  Compute importance score using Eq. 14
  for each candidate word w' in $N(w_i)$ **do**
    Compute replacement score using Eq. 15
  end for
  Select optimal replacement using Eq. 16
  if $S(w_i, w_i^*) > 0$ **then**
    Replace w_i with w_i^* in t
  end if
  if number of replaced words $\geq \epsilon_t$ **then**
    break
  end if
end for
return (v_T^{adv}, t^{adv})

B. Hyperparameter Analysis

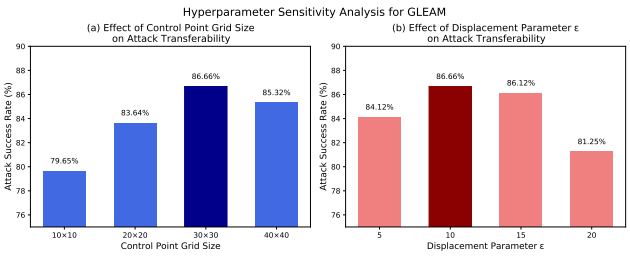

Figure 5. Hyperparameter sensitivity experiments for GLEAM. (a) Attack success rates (ASR) under different control point grid sizes (10×10 to 40×40) with fixed $\epsilon = 10$ when transferring from ALBEF to CLIP_{ViT} and TR models. (b) ASR under different displacement parameters ϵ (5, 10, 15, 20) with fixed 30×30 control point grid.

We conduct extensive experiments to determine the optimal configuration of key hyperparameters in our GLEAM framework. As illustrated in Fig. 5(a), the density of control points in the NURBS grid significantly impacts transferability. When transferring attacks from ALBEF to CLIP_{ViT} and TR models, a 30×30 control point grid achieves optimal performance with an 86.66% attack success rate (ASR). Sparser configurations (10×10, 20×20) provide insufficient local control for precise perturbations, resulting in lower transferability. Conversely, a denser grid (40×40) leads to slight performance degradation (85.32%), potentially due to over-parameterization of the transformation space.

For the displacement parameter ϵ (Fig. 5(b)), we observe that moderate values yield optimal results. With a fixed 30×30 control point grid, $\epsilon = 10$ achieves the highest ASR of 86.66%. Smaller displacements ($\epsilon = 5$) generate insufficient perturbations to effectively mislead target models, while larger values ($\epsilon \geq 15$) can disrupt semantic integrity and structural coherence, reducing transferability across different model architectures.

The global scaling factor r demonstrated relatively stable performance within the range [1.1, 1.8], which we adopted for all experiments. This range provides sufficient global distribution variation without compromising the structural integrity of the visual content. Based on these findings, we configure GLEAM with a 30×30 control point grid, displacement parameter $\epsilon = 10$, and scaling factor $r \sim U(1.1, 1.8)$ for optimal transferability.

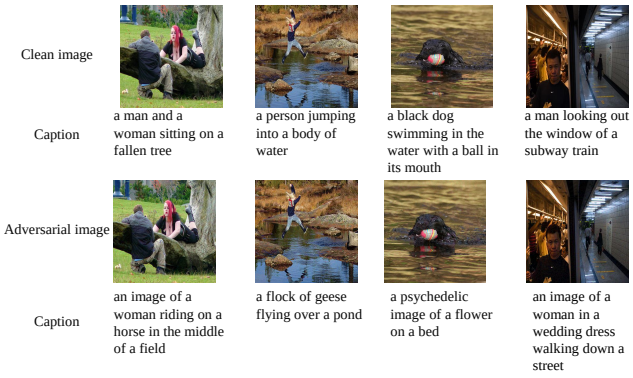

Figure 6. Visualization on Image Captioning Task. We use the ALBEF model, pre-trained on Image Text Retrieval (ITR) task, to generate adversarial images on the MSCOCO dataset and use the BLIP [16] model for Image Captioning on both clean images and adversarial images, respectively.

C. Qualitative Visualizations of Visual Grounding and Image Captioning

Fig. 6 presents qualitative results illustrating the effect of our GLEAM adversarial attack on image captioning performance. We first use the ALBEF model pre-trained on the Image-Text Retrieval (ITR) task to generate adversarial images from the MSCOCO dataset. We then feed both

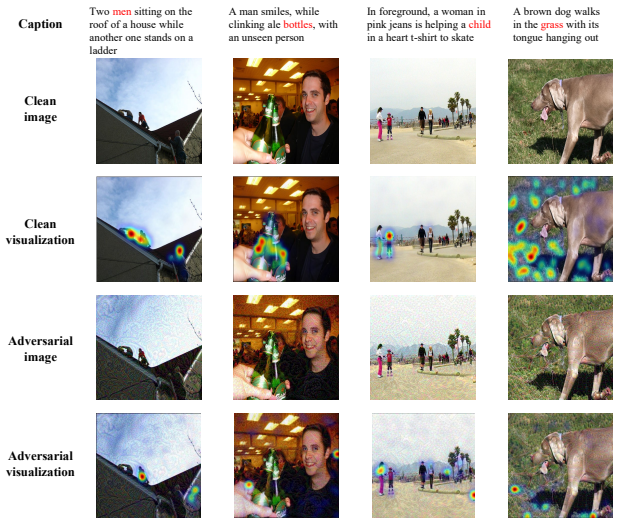

Figure 7. Visualization on Visual Grounding Task. We use the ALBEF model, pre-trained on the ITR task, to generate adversarial images on the RefCOCO+ dataset and use the same model, pre-trained on Visual Grouding (VG) task, to localize the regions corresponding to red words on both clean images and adversarial images, respectively.

the original clean images and their corresponding adversarial versions to the BLIP-2 model [16] for captioning. The results demonstrate that our adversarial examples significantly disrupt the image captioning process. These examples highlight GLEAM's effectiveness in generating transferable adversarial examples that successfully attack downstream tasks beyond the original ITR objective. The adversarial images maintain their visual appearance while significantly altering the feature representations used by the vision-language model for caption generation.

Fig. 7 demonstrates the impact of our GLEAM attack on the Visual Grounding task. Here, we generate adversarial images using the ALBEF model pre-trained on ITR and then evaluate these images using the same model architecture but pre-trained on the Visual Grounding task. The figure shows attention maps that highlight the image regions corresponding to the text phrases highlighted in red.

The comparison between clean and adversarial images reveals that our attack successfully disrupts the model's ability to correctly localize objects based on textual descriptions. In the clean images, the attention maps precisely highlight the objects mentioned in the text queries. However, in the adversarial versions, we observe: (1) attention drift, where the focus shifts to incorrect regions; (2) attention diffusion, where the attention becomes more scattered rather than concentrated on the target object; and (3) complete attention failure, where the model fails to identify any relevant region.

These results further validate GLEAM's cross-task transferability, showing that adversarial examples generated using the ITR objective effectively transfer to the Visual Grounding task. This cross-task attack effectiveness stems from our method's ability to perturb fundamental visual features that are shared across multiple vision-language tasks, demonstrating that GLEAM targets cross-modal alignment mechanisms common to various VLP models and tasks.