

433 A Samples for Training Dataset

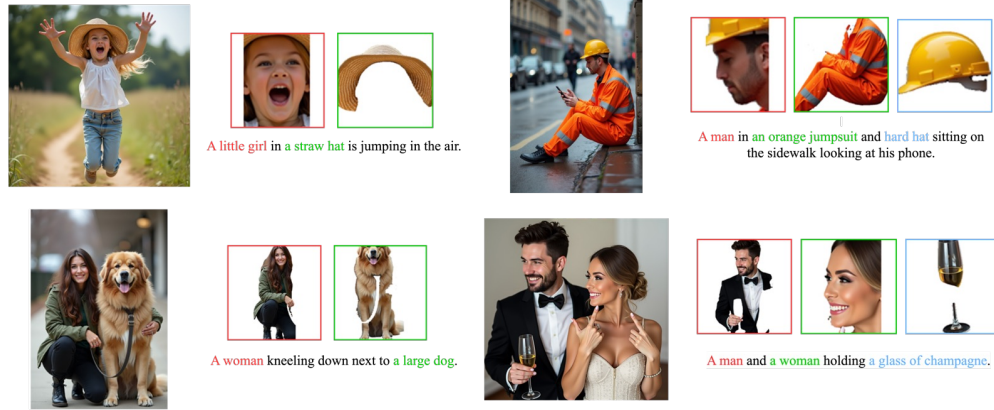


Figure 8: Examples of training data for multi-subject controlled generation.

Figure 8 presents examples of our training data for multi-subject controlled generation. As illustrated, the dataset covers a diverse range of scenarios, including human-object interactions, human-animal compositions, and complex multi-person scenes. For human-centric data, we intentionally randomly select facial images or full-body images as reference inputs. This strategy can further enhance the model’s generalization performance. By utilizing this diverse and extensive dataset, which encompasses a wide range of scene variations and control types, our model is able to achieve impressive editing capabilities while maintaining high consistency with the reference images.

441 B Impact of Prompt Variation on the Generated Image

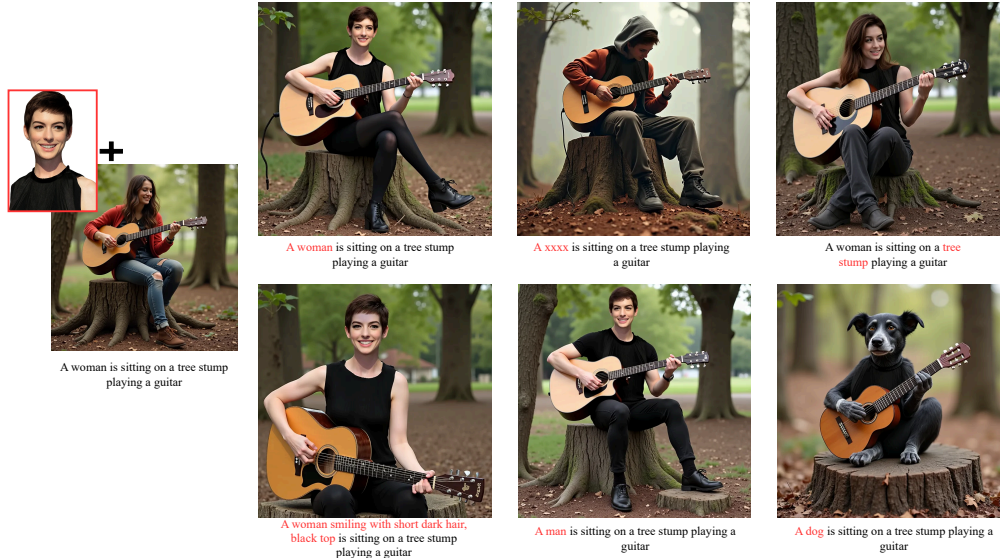


Figure 9: Impact of prompt variation on subject-controlled Image generation. The reference image and the initial output of our text-to-image generation model are shown on the left side. The right side illustrates the influence of different prompts on the generated output, with the prompt variances highlighted in red.

To evaluate the impact of prompt variation on subject-controlled image generation, we modified the injected words in the per-token text-modulation module while keeping the reference image constant. The results, shown in Figure 9, offer valuable insights. This visualization effectively illustrates

that a more detailed prompt description improves the preservation of the subject’s identity in the generated image. Additionally, when the prompt closely matches the reference image, our model not only incorporates intricate image details but also maintains a high level of control over the subject’s attributes, even allowing for successful changes such as gender. On the other hand, if there is a significant semantic mismatch between the injected prompt and the reference image (e.g., trying to generate a person’s image from prompts like “a dog” or “a tree stump”), the injection process consistently fails. This highlights our model’s ability to accurately target and incorporate reference image features into specific words, enabling precise control over the generated output.

C Comparison of the CLIP-T and DPG scores

When evaluating text-to-image generation models, the CLIP-T [24] score has been a prevalent metric in prior studies, assessing semantic consistency by leveraging CLIP’s image-text embeddings. However, our research highlights the superior efficacy of the DPG (Dense Prompt Graph) score, particularly for intricate prompts. While CLIP-T offers a broad measure of semantic alignment, the DPG score is specifically designed to evaluate a model’s capacity to interpret and execute detailed and complex textual instructions. It rigorously assesses editing abilities across multiple objects, diverse attributes, and intricate relationships, thereby capturing the nuanced and fine-grained semantic alignment crucial for advanced compositional generation. This provides a more comprehensive and robust evaluation for challenging scenarios.

D Illustration of Region Preservation Loss

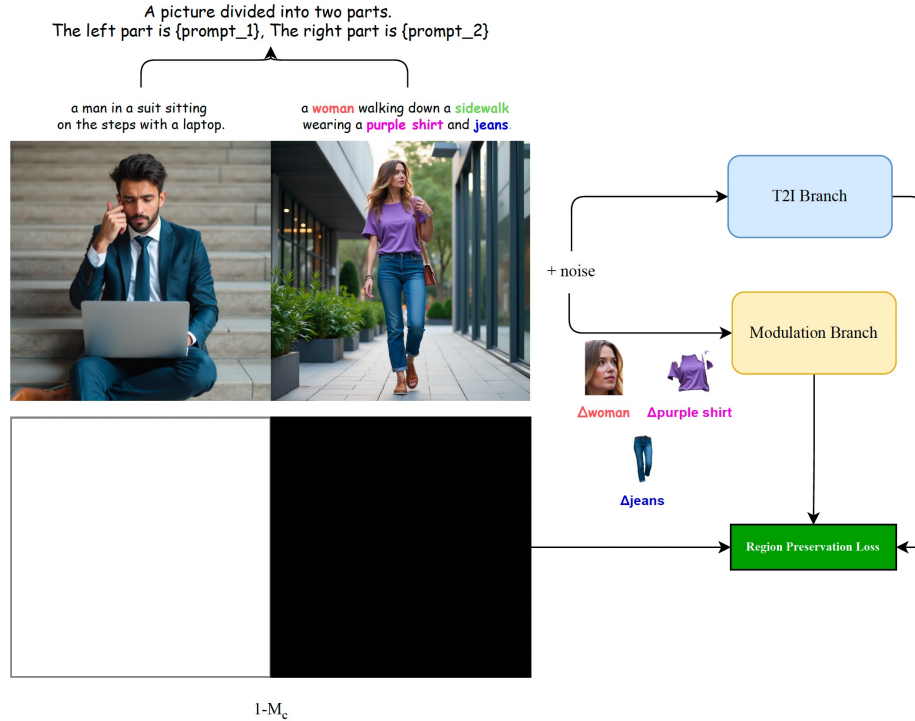


Figure 10: Illustration of the region preservation loss.

Figure 10 shows the illustration of our region preservation loss. We form training samples by concatenating two existing samples, merging their captions, and randomly applying modulation to only one side. For the unmodulated regions, defined by M_c , we enforce consistency between our model’s output ($V_{\theta'}(z_t, t, y^*)$) and the text-to-image branch’s output ($V_{\theta}(z_t, t, y)$) via an L2 loss (Eq. 1). By using this regularization, XVerse can better inject the reference image into specific areas without affecting the generation of irrelevant areas, thereby achieving more precise generation control.

471 E Ablation Study for Text-Image Attention Loss

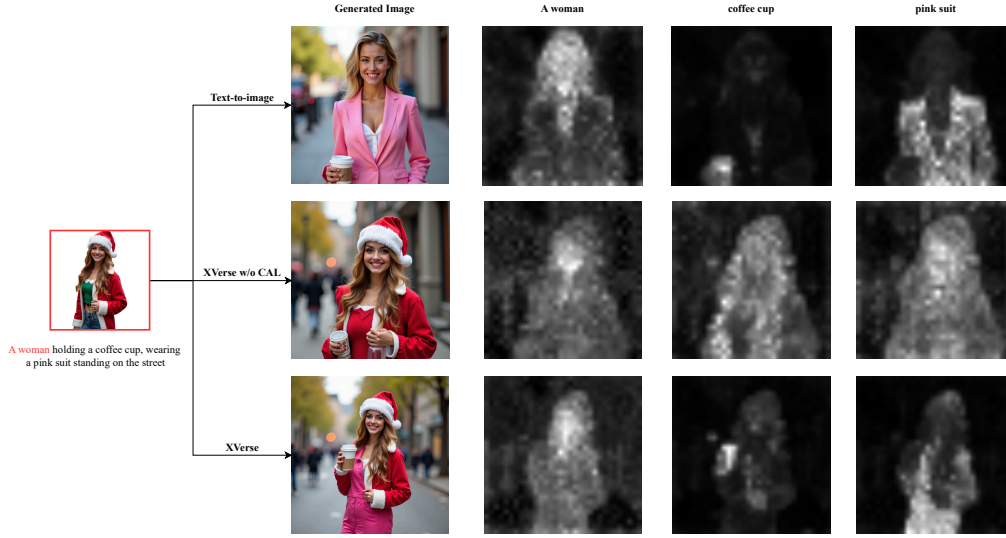


Figure 11: The qualitative comparison of Text-Image Attention Loss. This image shows the generated results and attention maps for "woman", "coffee cup", and "pink suit" for each method.

472 To validate the effectiveness of our text-image attention loss, we conducted an experiment where we
 473 excluded this regularization and examined the generated outputs along with their respective attention
 474 maps. The qualitative analysis presented in Figure 11 clearly demonstrates the significance of this
 475 method. It demonstrates our method's ability to maintain the structural and editable characteristics
 476 of the T2I branch following modulation injection. Through ensuring L2 consistency between the
 477 cross-attention maps of the modulated model and the reference T2I branch, our approach ensures the
 478 reliable preservation of text-image semantic interactions. This ultimately enables precise control over
 479 semantics, as visually evidenced by the generated results and attention maps for specific prompts like
 480 "woman," "coffee cup," and "pink suit."

481 F Broader Impacts

482 Our model, XVerse, marks a significant leap in multi-subject controllable text-to-image generation,
 483 leading to enhanced fidelity and editability. This breakthrough holds substantial positive societal
 484 impacts, particularly within the creative industries, where it can revolutionize the creation of person-
 485 alized and complex visual content. Furthermore, XVerse can transform education and training by
 486 providing more engaging and tailored visual aids, and contribute to content inclusivity by enabling
 487 the representation of a wider range of individuals and scenarios.

488 However, this powerful technology also presents potential negative societal impacts. The improved
 489 generation capability could lead to misinformation and deepfakes, raise privacy concerns if used
 490 improperly, and potentially amplify biases present in training data. As foundational research, XVerse
 491 isn't directly tied to deployment. Yet, we believe it's crucial to acknowledge these risks. Future work
 492 will explore mitigation strategies like content detection and ethical guidelines, contributing to the
 493 responsible advancement of generative AI.