A APPENDIX

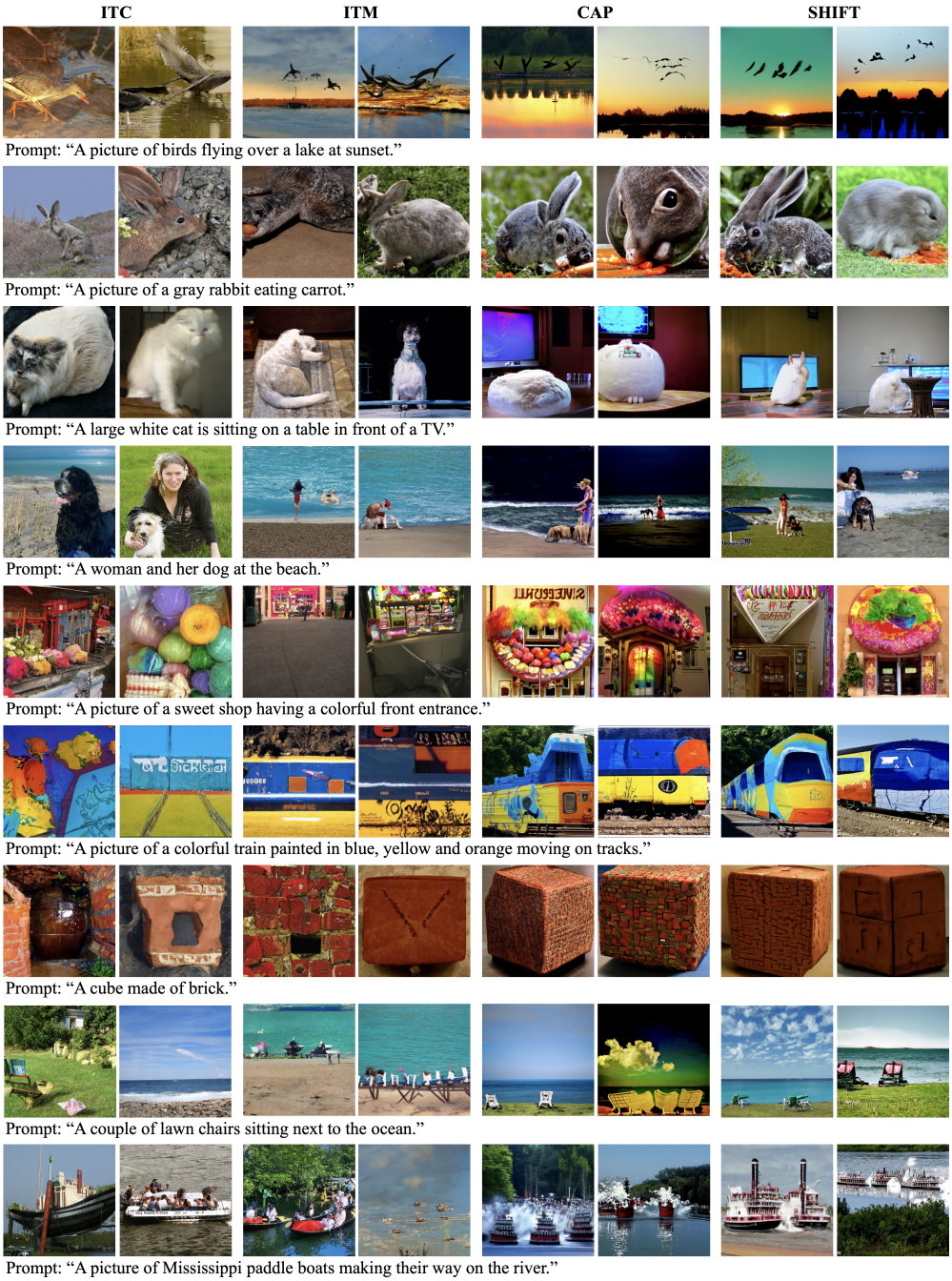

Figure 4: Additional text-to-image generation examples. We can consistently observe that while **ITC** focuses on detailed formulation of the salient object, **CAP** and its variant **SHIFT** understand the prompt in a finer level and output more faithful visualizations. It is also apparent that **ITC** alone often leaves out certain objects or mixes different visual semantics (*e.g., colorful entrance*).

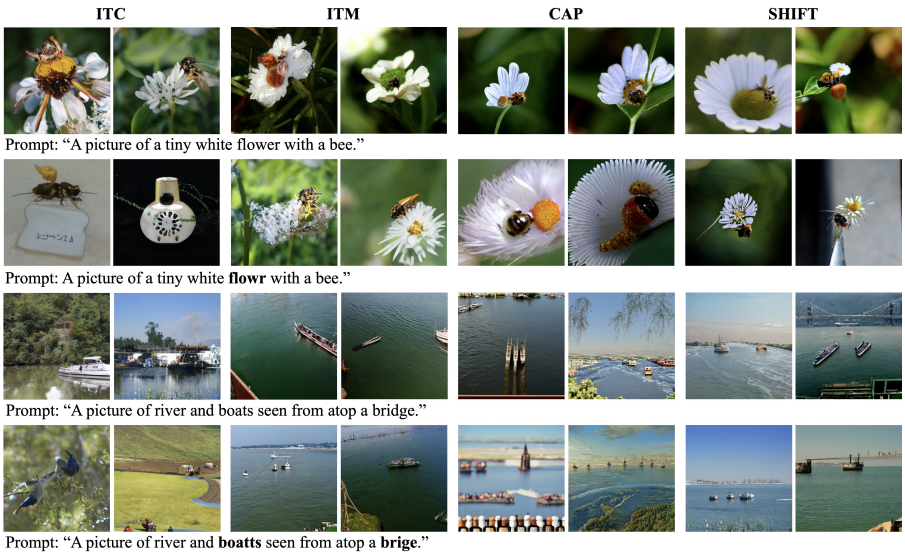

Figure 5: Additional experiments for noise robustness. Although **ITC** produces realistic images with clean prompts, minor typos can completely ruin their semantic signals. In contrast, losses that provide denser supervisions generally output consistent results despite textual noise, showing better robustness.

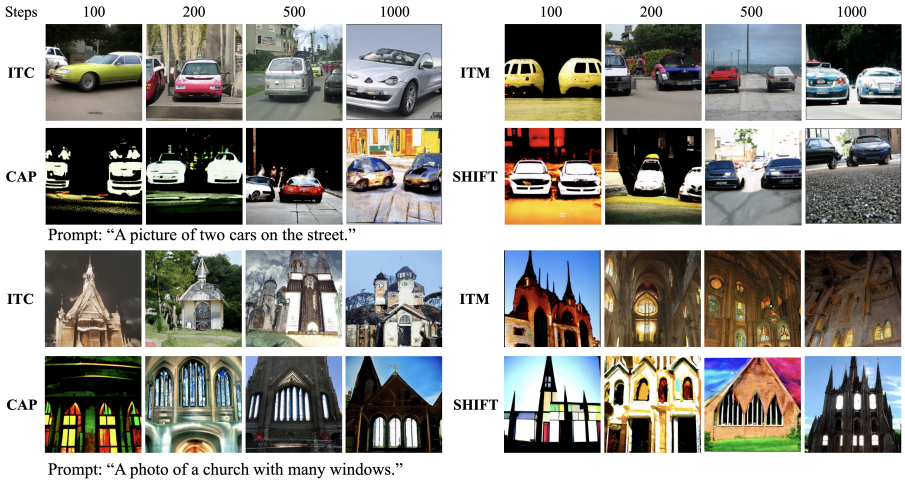

Figure 6: Additional results for optimization complexity. Captioning-based losses require more diffusion steps to generate realistic images, while **ITC** and **ITM** quickly forms reasonable shapes and appearances.

Figure 7: Qualitative comparison between baselines that combine multiple objectives. **BLEND** mixes **CAP** and **ITC** with no transition. We observe that gradually shifting from **CAP** to **ITC** enjoys advantages from both sides, *i.e.,* faithful scene composition and realistic details.