A. appendix

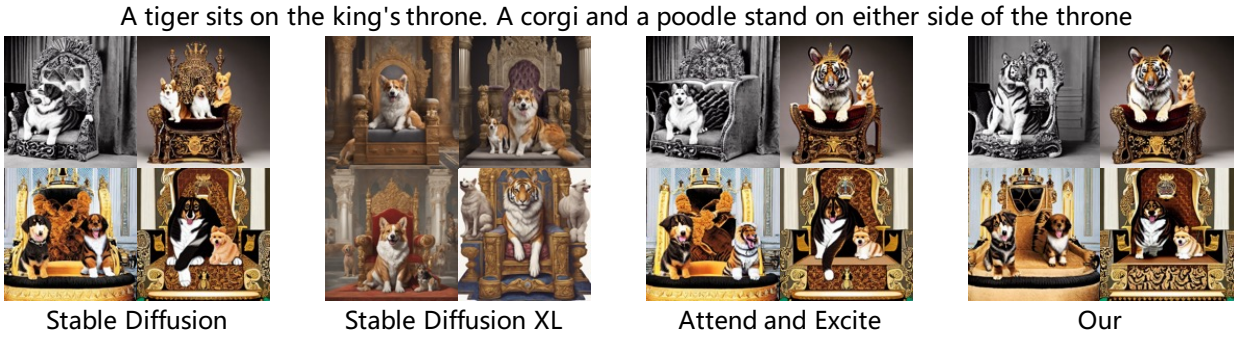

Figure 1. It's a challenging scenario; both methods struggle to generate faithful images in complex scenes featuring multiple entities.

Table 1. Human evaluation. Here, we have two types of datasets. The first type, ALL, comprises two kinds of prompts: one featuring [entity A] and entity B, and another featuring [attribute A] [entity A] and [attribute B] [entity B]. The prompts containing attributes constitute only one-third of the total. The second type, Binding, exclusively includes the latter kind of prompt. We both selected 12 prompts, each capable of generating 64 images.

MODEL	ALL	BINDING
STABLE DIFFUSION	1.81	2.09
LINGUISTIC BINDING	17.36	**31.06**
ATTEND AND EXCITE	18.04	12.42
OURS	**35.23**	25.09
NO MAJORITY WINNER	27.56	29.34

Table 2. We have mainly used prompts of the following form: a [entity A] and [entity B].

MODEL	FID
STABLE DIFFUSION	**84.873**
OURS	89.671

Figure 2. More case.