Appendix

A. Dataset Details

We select 42 videos from the DAVIS dataset [34], covering a range of categories including animals, vehicles, and humans. The selected video items are listed in Tab. 2. To obtain video footage, we use BLIP-2 [23] for automated captions. We then manually design three edited prompts for each video, resulting 140 edited prompts in total. These edited prompts include object editing, background changes, and style transfers, as described in Sec. 4.

Table 2: ***Names of videos selected from DAVIS dataset.***

bear	blackswan	boat
breakdance-flare	camel	car-roundabout
car-shadow	car-turn	cows
dog	dog-agility	drift-turn
elephant	flamingo	girl-dog
gold-fish	golf	guitar-violin
hike	hockey	horsejump-high
horsejump-low	kid-football	kite-surf
lab-coat	libby	lions
longboard	lucia	mallard-water
man-bike	mbike-santa	mbike-trick
motorbike	parkour	rhino
running	scooter-gray	snowboard
swing	tandem	tennis

B. User Study Details

We conduct a user study on our dataset of 140 edited prompts to compare our method against two baselines: Plug-and-Play [45] and CogVideo [20]. The comparison results are shown in Tab. 1. The participants of the user study are mainly students and colleagues in university. We ask 5 raters to evaluate each edited prompt by comparing two videos generated by two different methods (shown in random order) and answering two following questions:

1. Which video has higher consistency? Please select the one that looks more smooth as a video.
2. Which video matches the text better? Please select the one that better represents the given text description.

C. Additional Results

Fig. 10 and Fig. 11 showcase additional video examples of our methods, Fig. 12 provides additional comparison with baselines, and Fig. 13 gives additional results of ablation study.

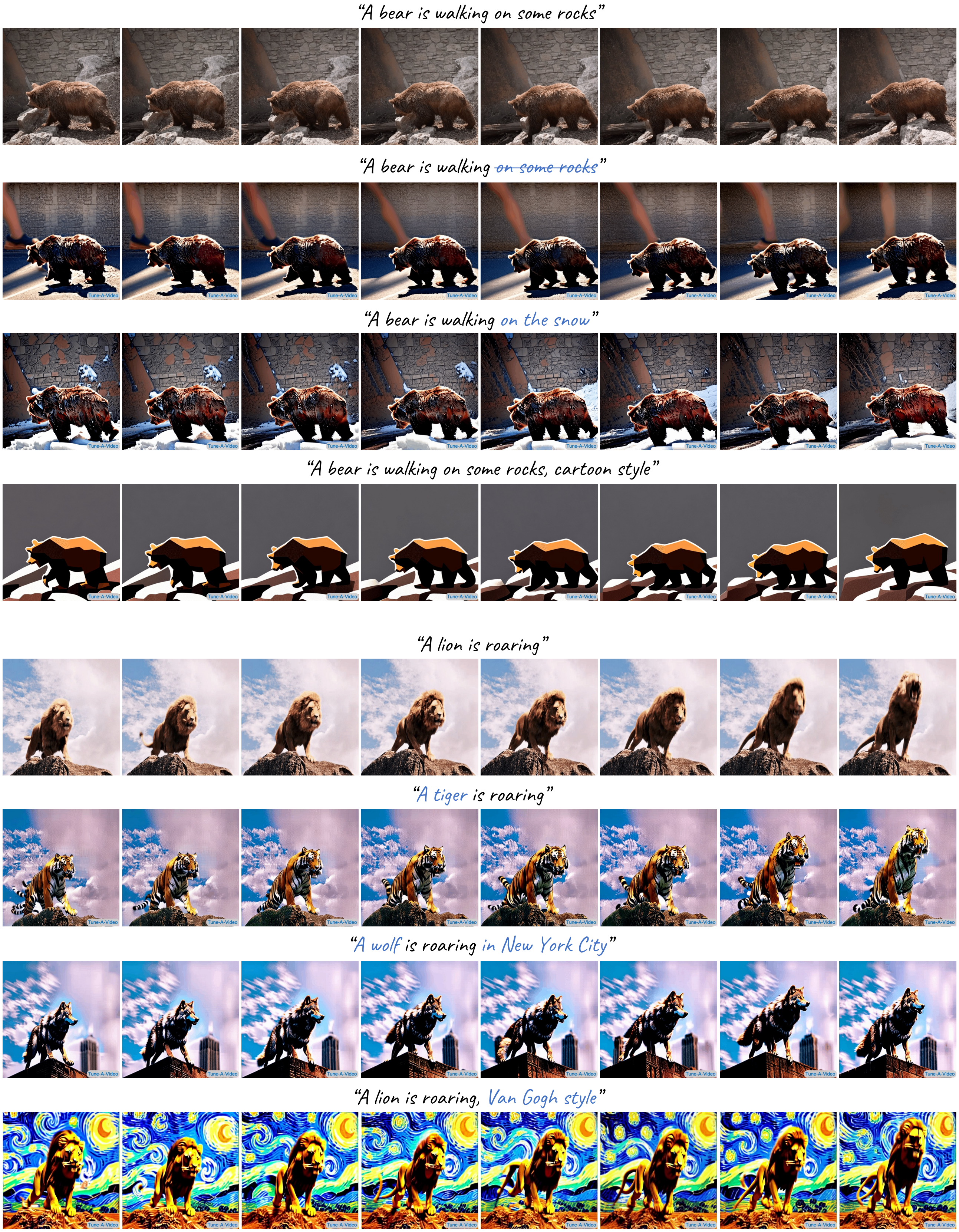
"A bear is walking on some rocks"
"A bear is walking ~~on some rocks~~"
"A bear is walking on the snow"
"A bear is walking on some rocks, cartoon style"
"A lion is roaring"
"A tiger is roaring"
"A wolf is roaring in New York City"
"A lion is roaring, Van Gogh style"
Tune-A-Video

Figure 10: ***Additional sample results of our method (1/2).***

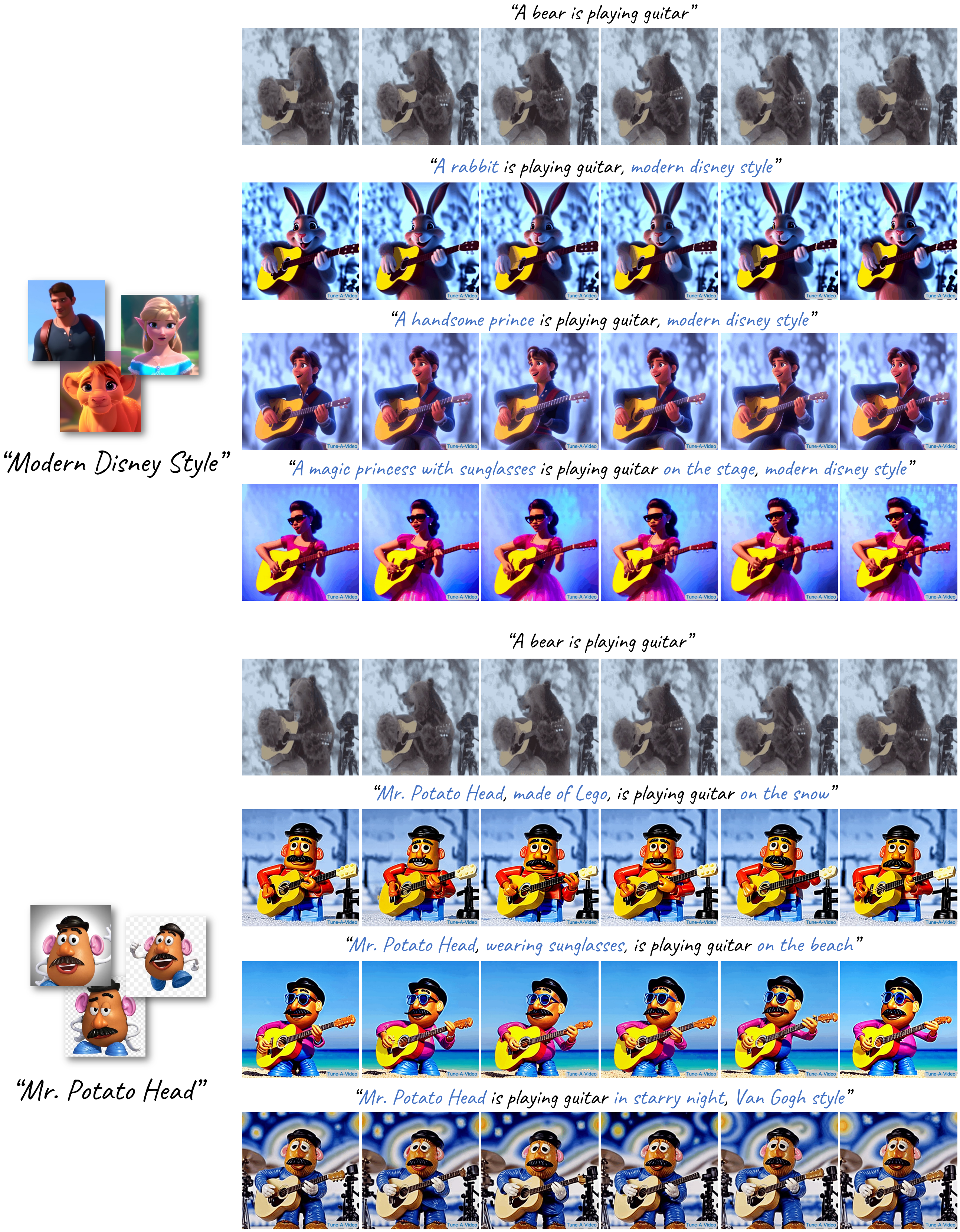
"A bear is playing guitar"
"A rabbit is playing guitar, modern disney style"
"A handsome prince is playing guitar, modern disney style"
"Modern Disney Style"
"A magic princess with sunglasses is playing guitar on the stage, modern disney style"
"A bear is playing guitar"
"Mr. Potato Head, made of Lego, is playing guitar on the snow"
"Mr. Potato Head, wearing sunglasses, is playing guitar on the beach"
"Mr. Potato Head"
"Mr. Potato Head is playing guitar in starry night, Van Gogh style"

Figure 11: ***Additional sample results of our method (2/2).***

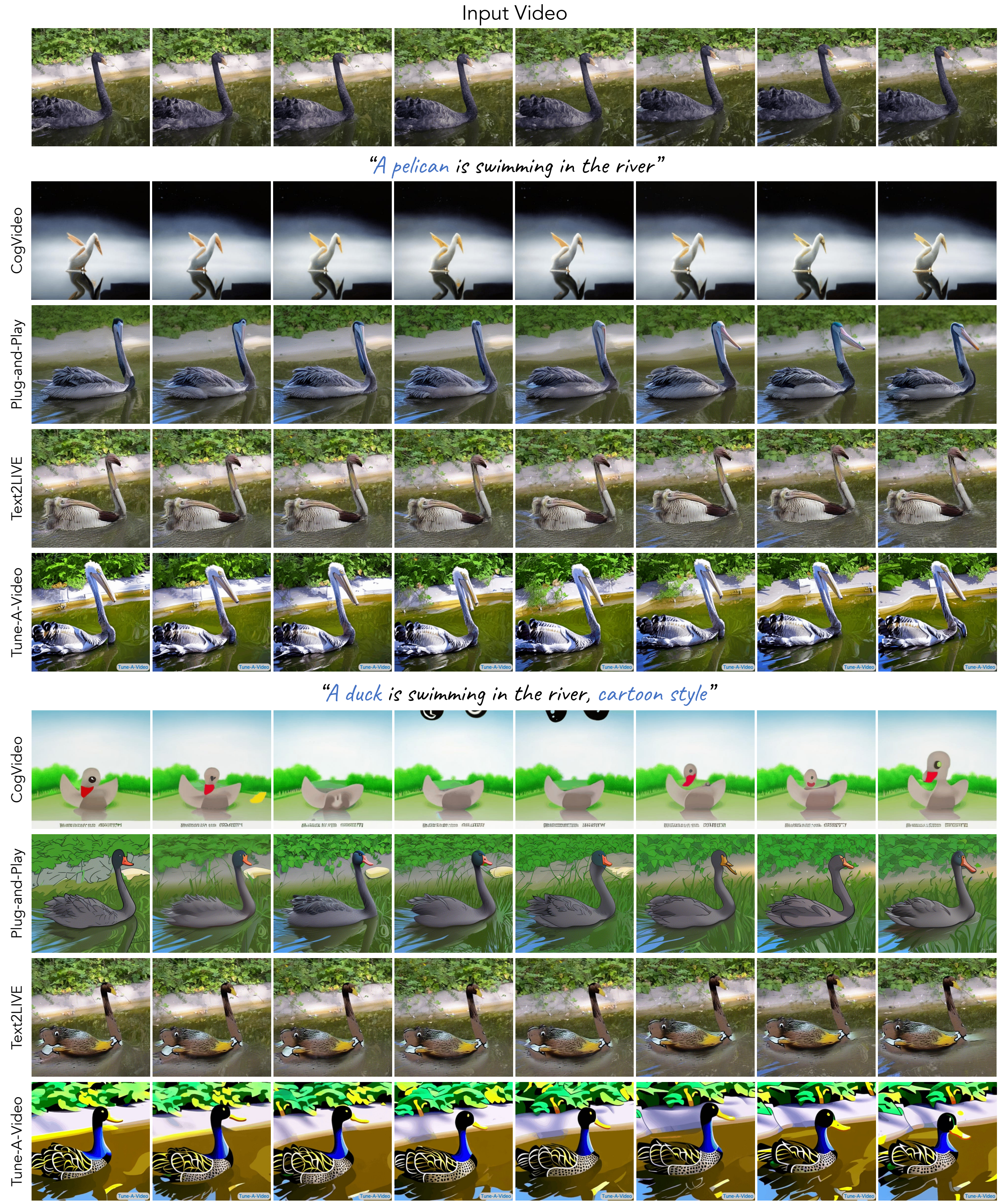

Figure 12: ***Additional qualitative comparsion between evaluated methods.***

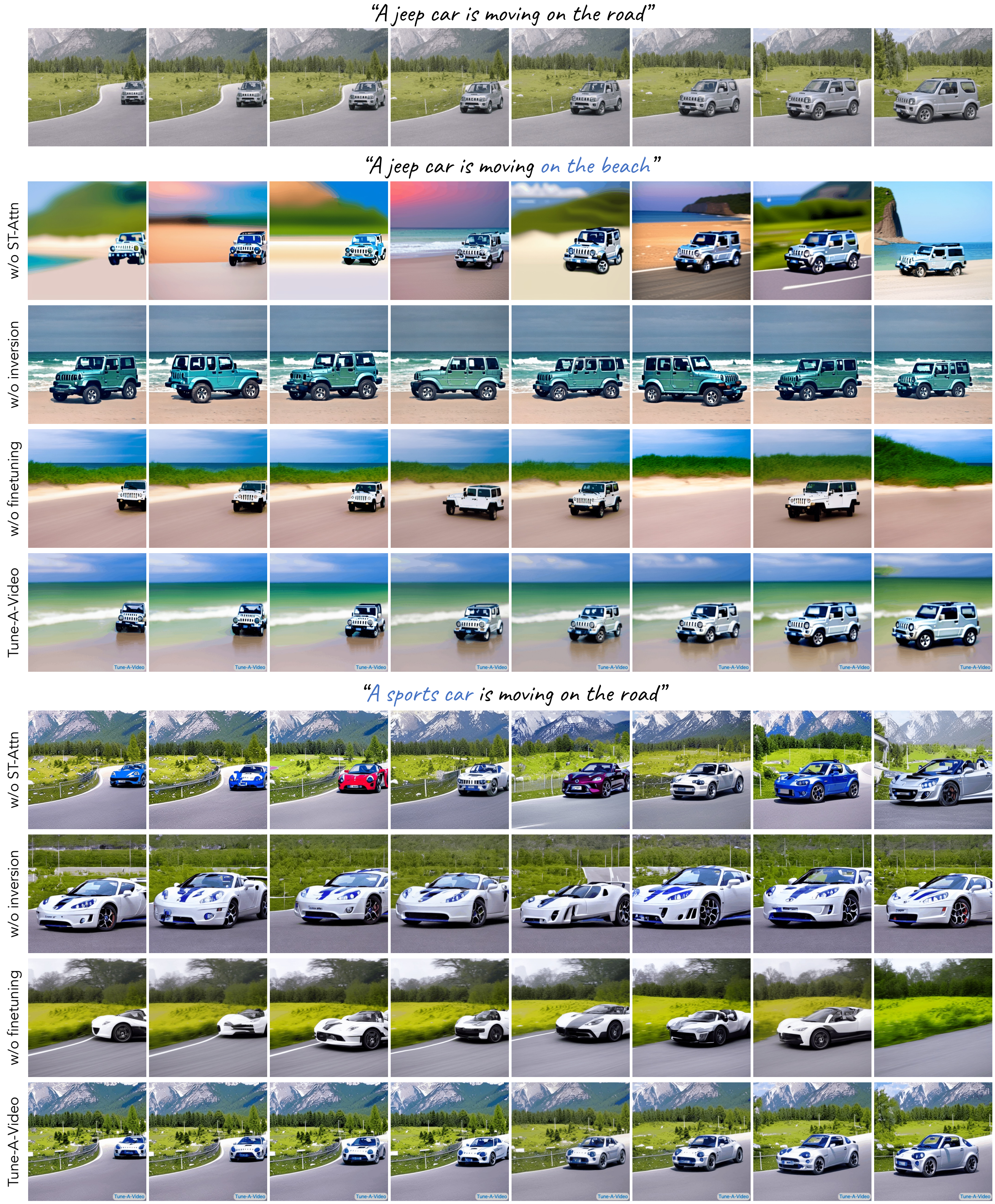
"A jeep car is moving on the road"
"A jeep car is moving on the beach"
w/o ST-Attn
w/o inversion
w/o finetuning
Tune-A-Video
"A sports car is moving on the road"
w/o ST-Attn
w/o inversion
w/o finetuning
Tune-A-Video

Figure 13: ***Additional ablation study.***