

NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion

Chenfei Wu^{1*} Jian Liang^{2*} Lei Ji¹ Fan Yang¹ Yuejian Fang^{2†}
Daxin Jiang¹ Nan Duan^{1†}

¹ Microsoft Research Asia ² Peking University

A wooden house sitting in a field.



A young girl eating a very tasty looking slice of pizza.



Walnuts are being cut on a wooden cutting board.



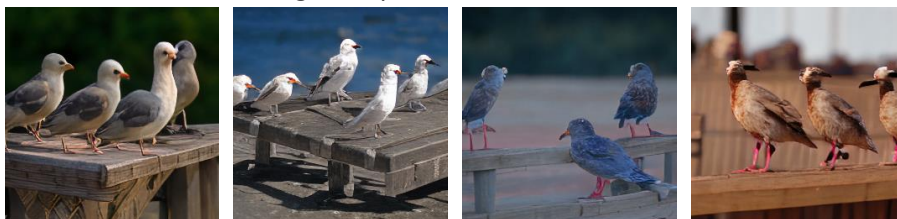
A boy with a hat wearing a tie.



Fig. 1: More samples of Text-to-Image (T2I) task generated by NÜWA.

* Both authors contributed equally to this research. † Corresponding author.

Some birds are standing on top of a wooden bench.



A dog wearing a Santa Claus hat is lying in bed.



A child is sitting in front of a cake with candles.



A bowl of food with meat in a sauce, broccoli and cucumbers.



Fig. 2: More samples of Text-to-Image (T2I) task generated by NÜWA.

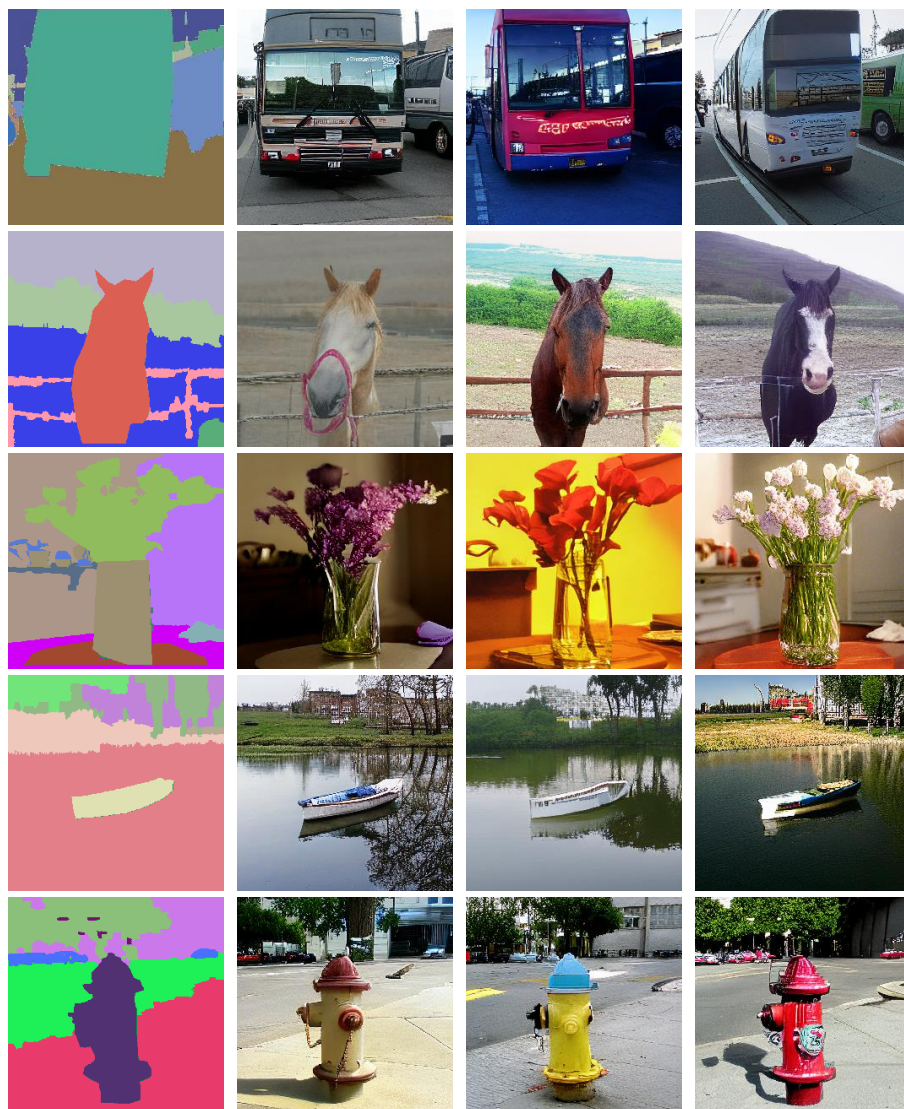


Fig. 3: More samples of Sketch-to-Image (S2I) task generated by NÜWA.

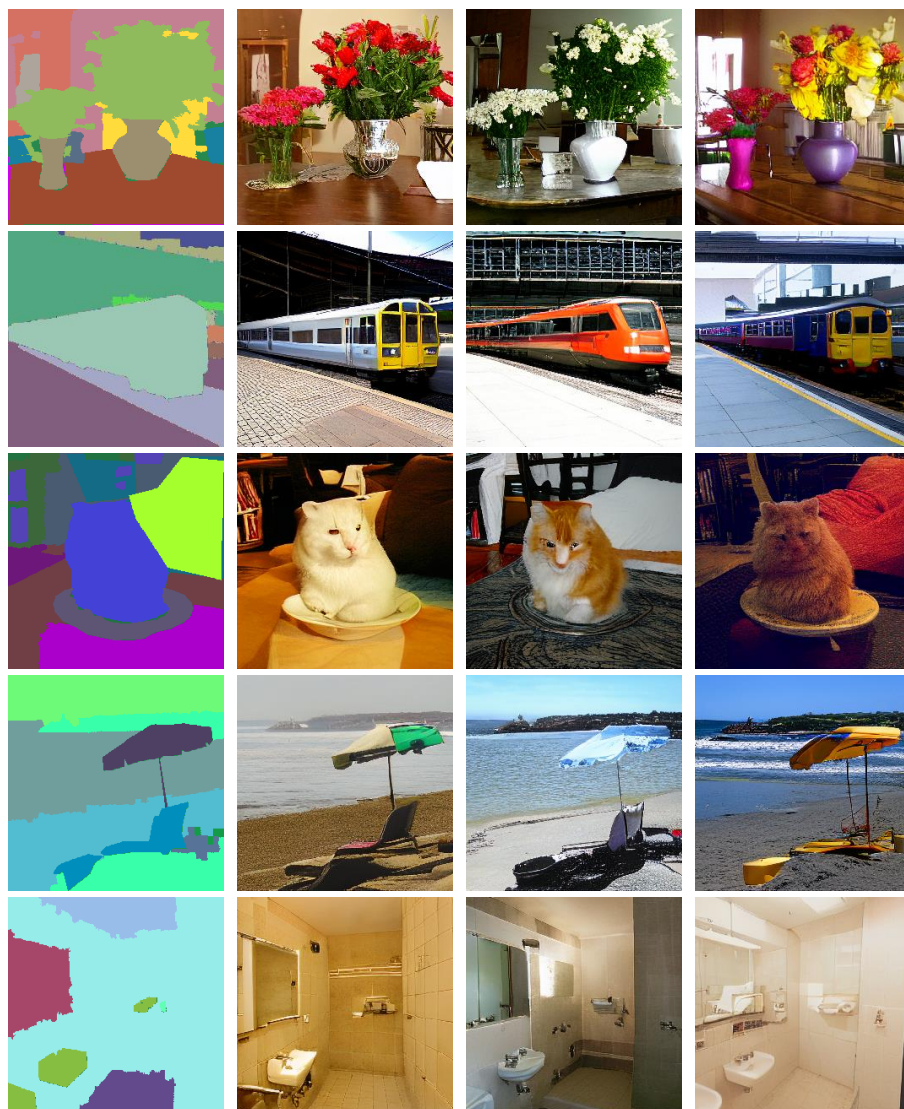


Fig. 4: More samples of Sketch-to-Image (S2I) task generated by NÜWA.

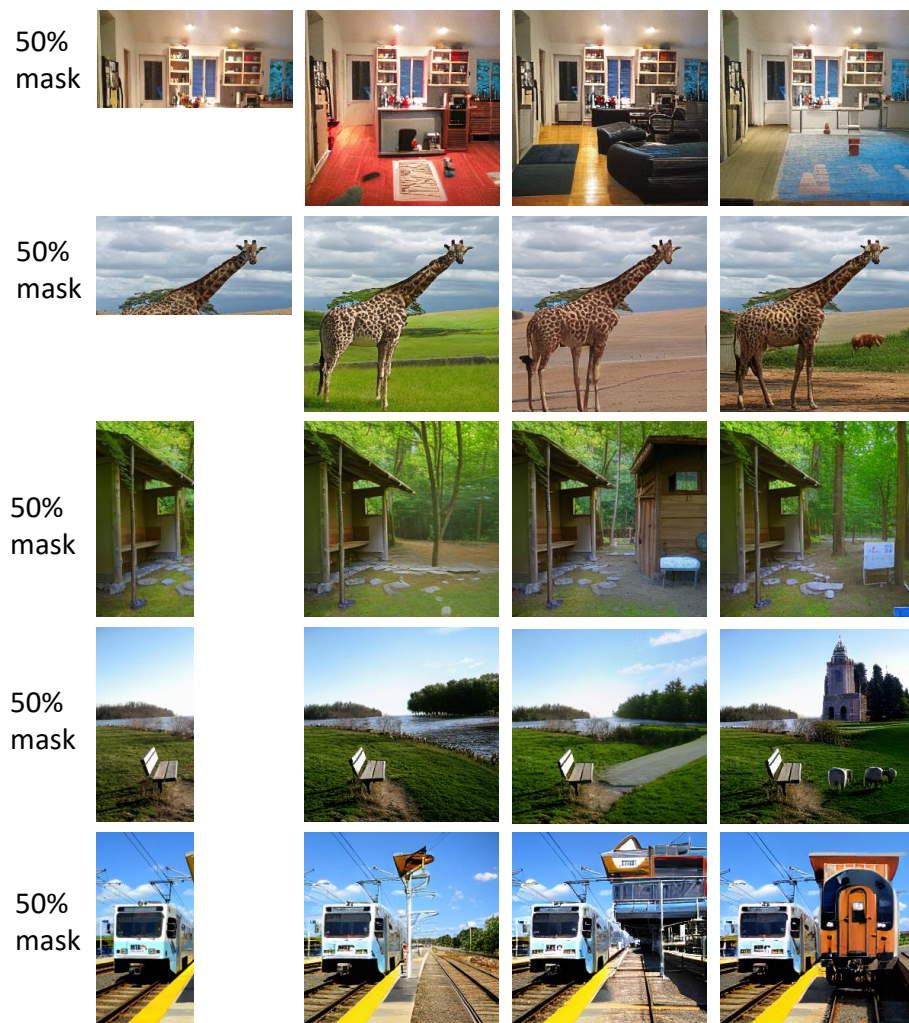


Fig. 5: More samples of the Image Completion (I2I) task generated by NÜWA.

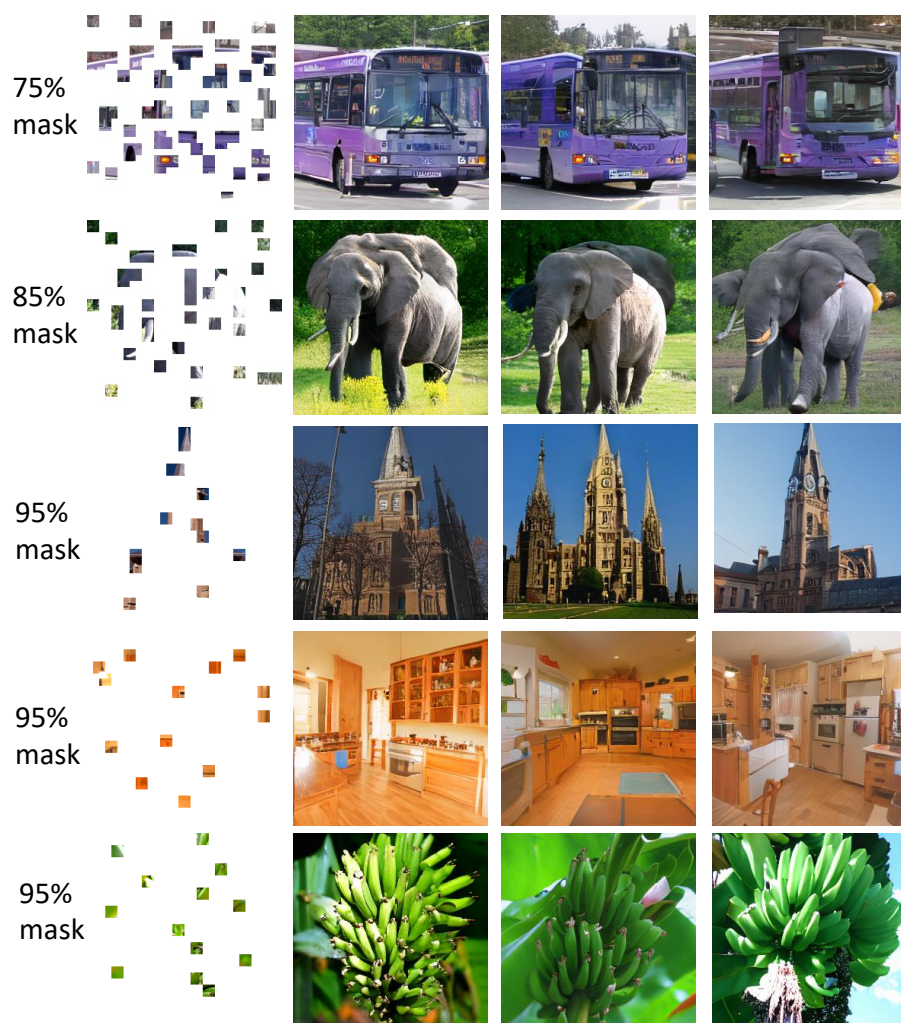


Fig. 6: More samples of the Image Completion (I2I) task generated by NÜWA.

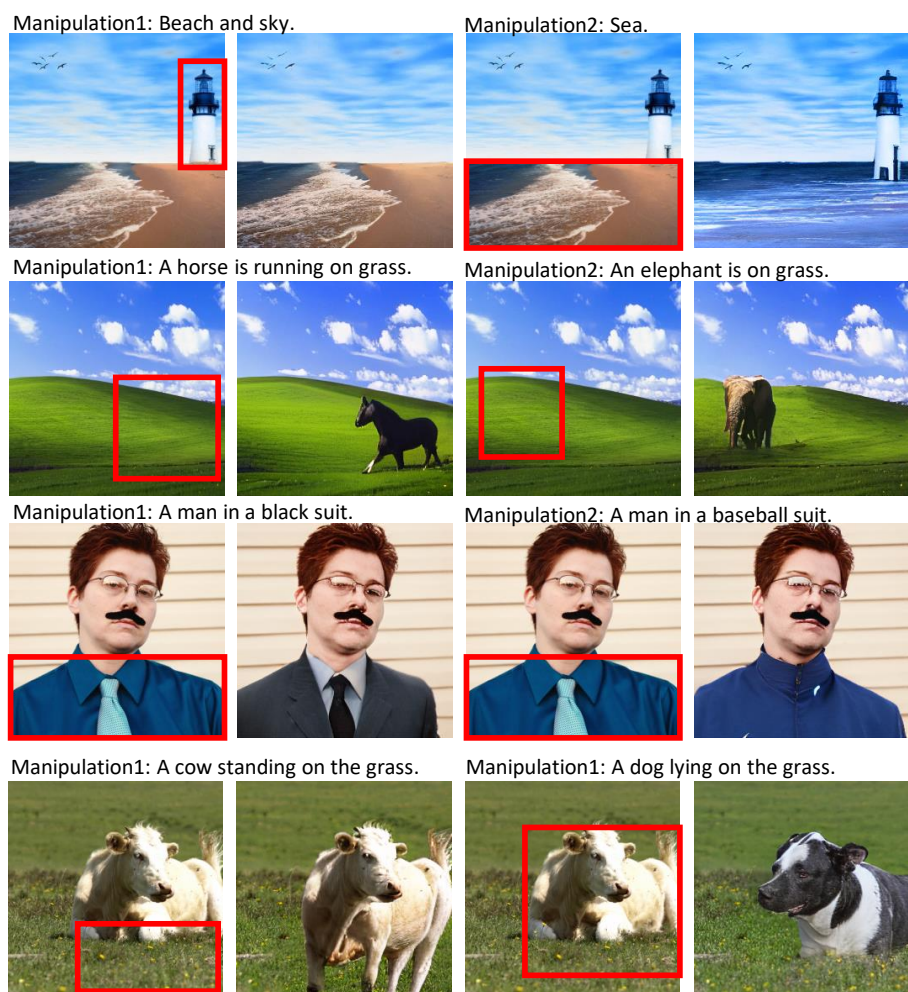


Fig. 7: More samples of the Text-Guided Image Manipulation(TI2I) task generated by NÜWA.

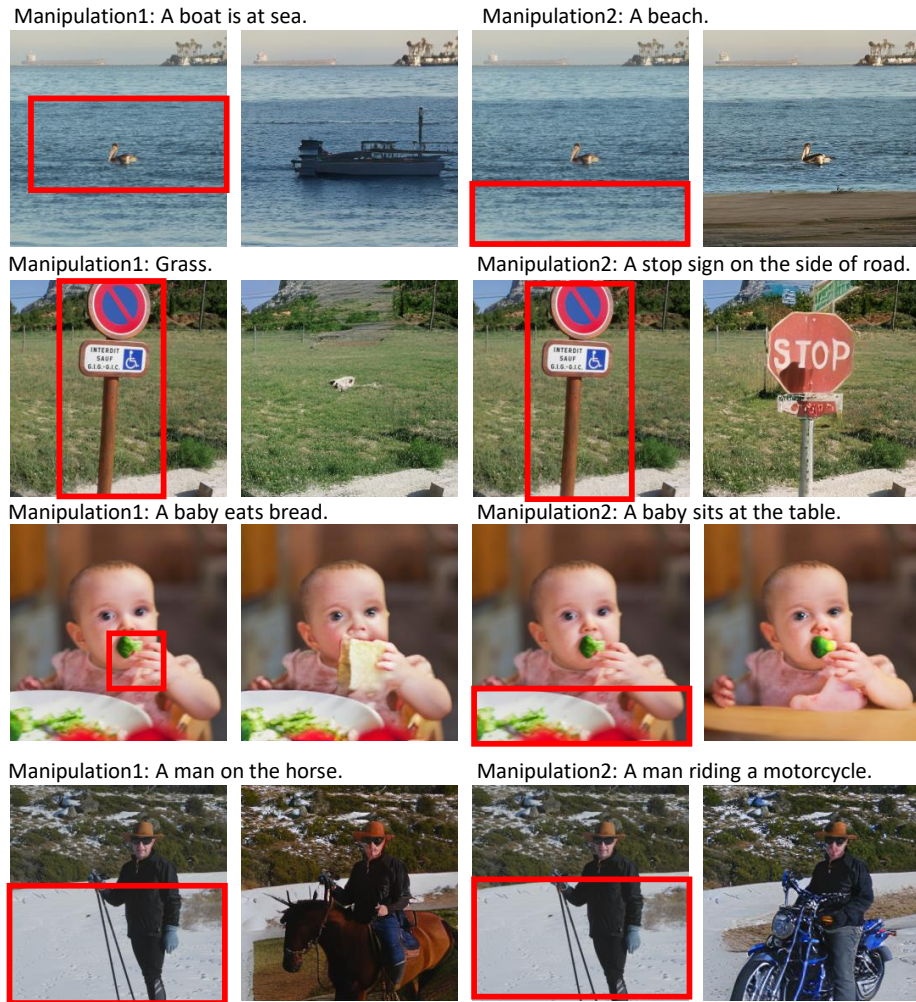
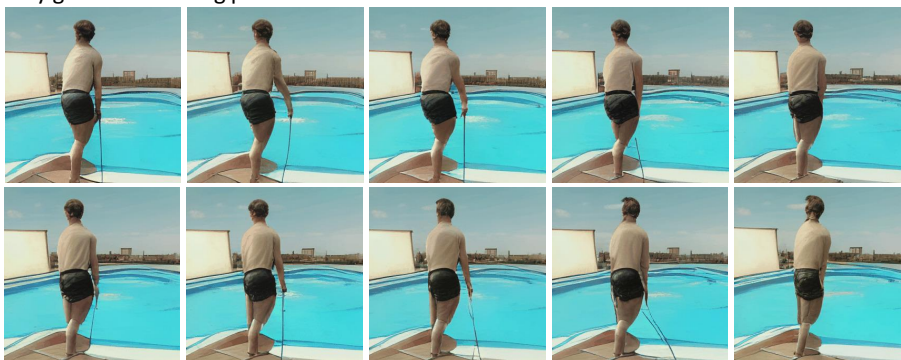


Fig. 8: More samples of the Text-Guided Image Manipulation(TI2I) task generated by NUWA.

Play golf on grass.



Play golf at swimming pool.



Running on the sea.

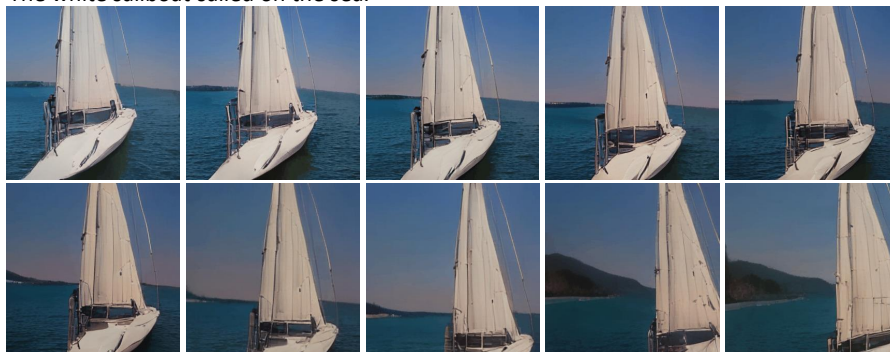


Fig. 9: More samples of the Text-to-Video (T2V) task generated by NÜWA.

A suit man is talking from a studio with fun.



The white sailboat sailed on the sea.



A man is folding a piece of yellow paper.

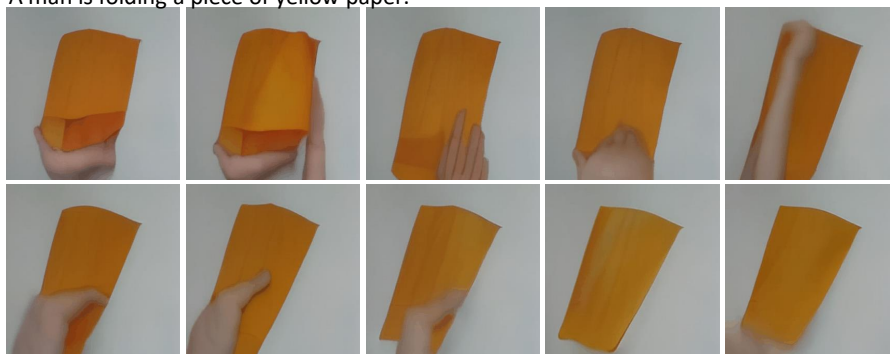


Fig. 10: More samples of the Text-to-Video (T2V) task generated by NÜWA.

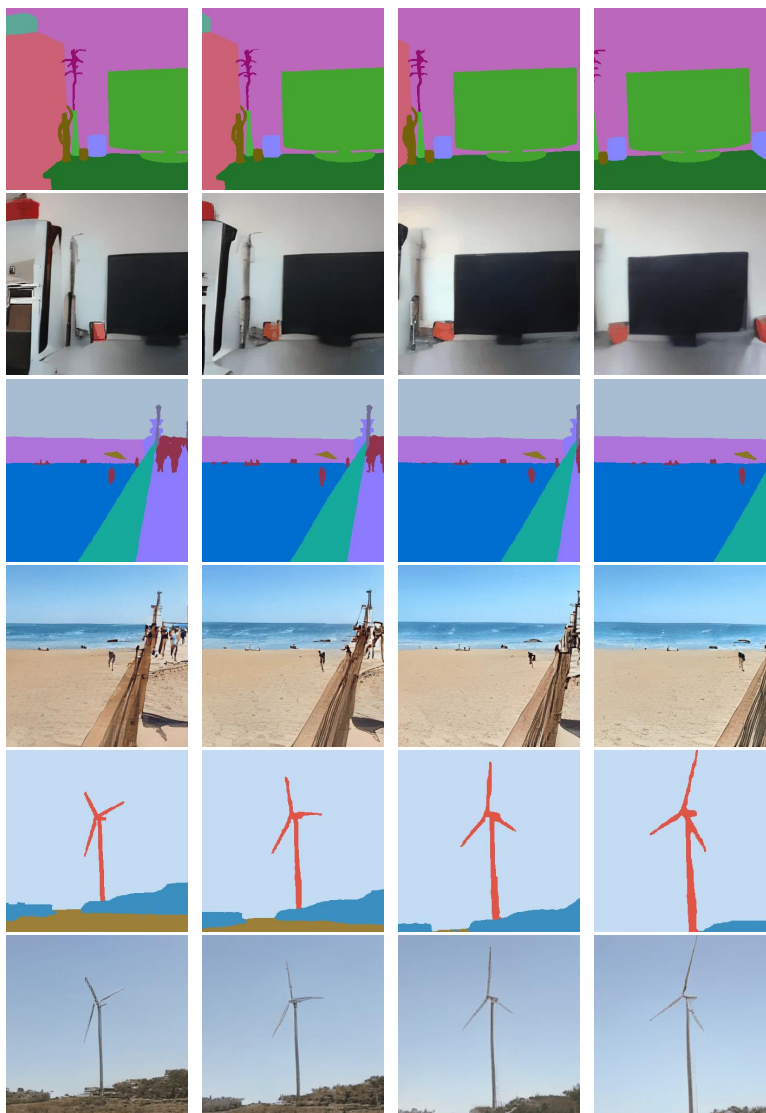


Fig. 11: More samples of Sketch-to-Video (S2V) task generated by NÜWA.

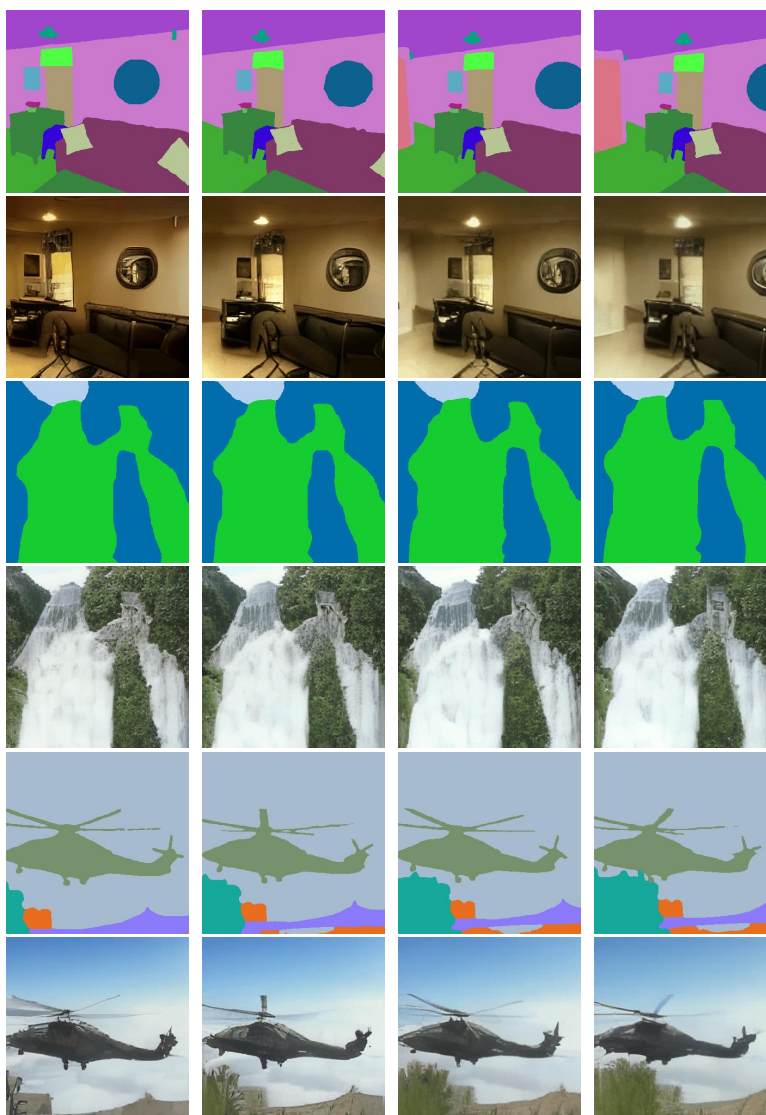


Fig. 12: More samples of Sketch-to-Video (S2V) task generated by NÜWA.



Fig. 13: More samples of the Video Prediction (V2V) task generated by NÜWA. Only one frame (see red box) is used as condition.

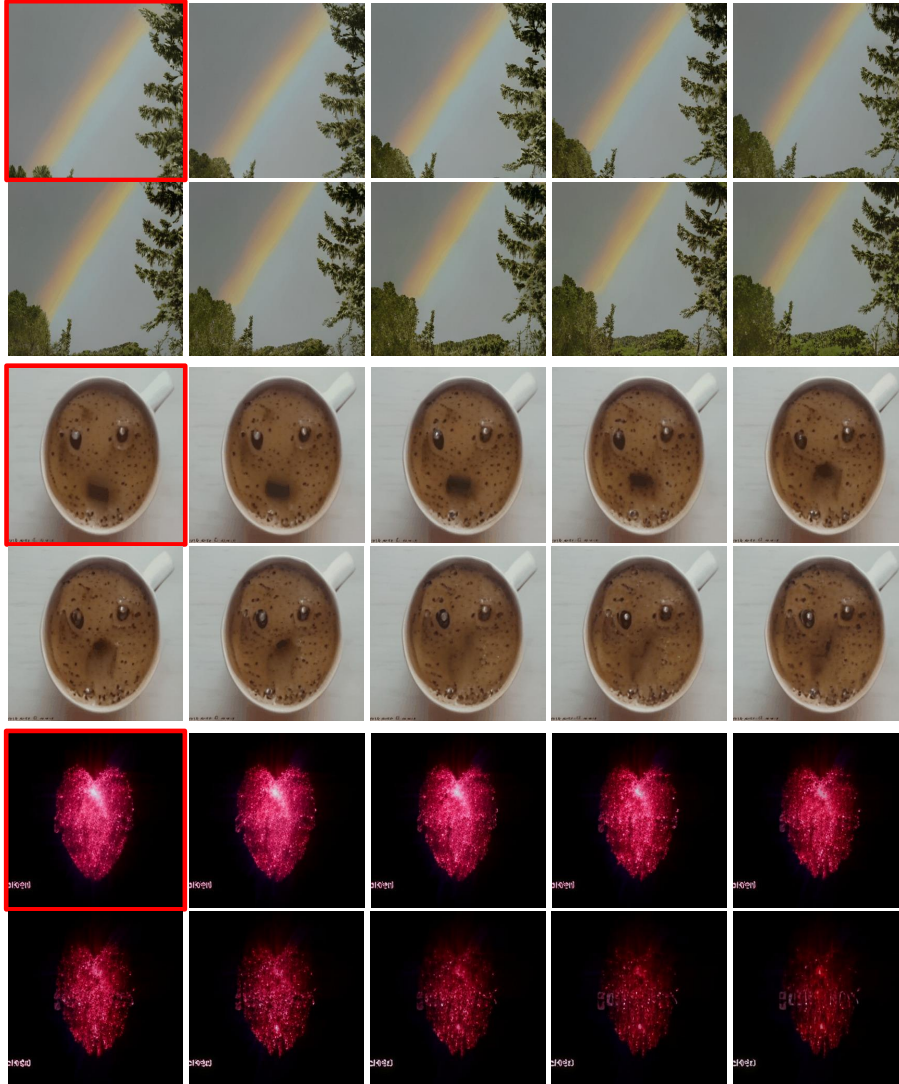
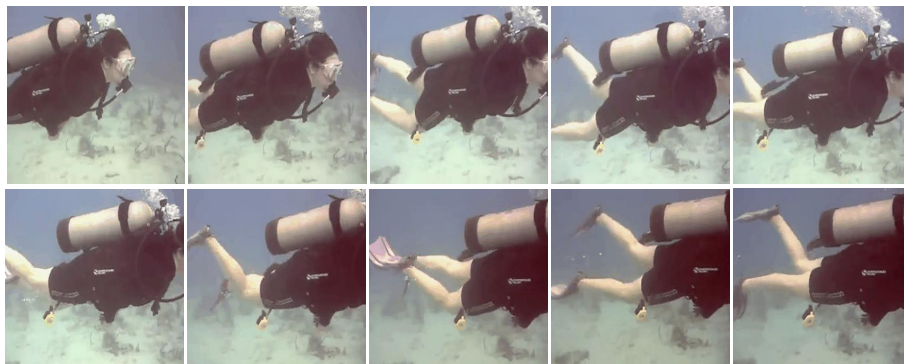


Fig. 14: More samples of the Video Prediction (V2V) task generated by NÜWA. Only one frame (see red box) is used as condition.

Raw Video:



Manipulation1: The diver is swimming to the surface.



Fig. 15: More samples of Text-Guided Video Manipulation (TV2V) task generated by NÜWA.

Manipulation2: The diver is swimming to the bottom.



Manipulation3: The diver is flying to the sky.



Fig. 16: More samples of Text-Guided Video Manipulation (TV2V) task generated by NUWA.