1 Appendix

In this appendix section, we provide following items:

- Sec. 1.1 Ablation experiments on the language model in our **OneDiff** framework and the hyper-parameters settings of the VDM module.
- Sec. 1.2 The statistics of our proposed **DiffCap Dataset**, both real-image and synthetic-image samples from DiffCap and some potential failure cases.

1.1 Ablation

This section further dig into the modules promoting high performances in OneDiff. Ablation study is conducted on the selection of language models and the selection of hyper-parameters for VDM modules.

Ablation on the Language Models. Considering the power brought by the large language model (LLM), we simply substitute the language model with Qwen2-1.5B to see the improvement that benefits from our framework. OneDiff still achieves considerably high performances on the Birds-to-Words benchmarks with lower computational costs.

Table 1: Ablation of model design on language models. We conduct ablation analysis on the Birds-to-Words benchmark.

LLM	B	C	M	R
Vicuna-7B	25.0	21.8	25.9	46.7
Qwen2-1.5B	23.8	20.1	25.1	46.7

Ablation on the Hyper-Parameters of VDM. The hyper-parameter setting of VDM is explored in Tab. 2. For efficiency, we only train 3000 steps in Stage II when exploring the model design. The number of delta tokens varies in 16, 32, 64 and 128 with 6 or 12 Transformer layers. OneDiff is not quite sensitive to different hyper-parameters with relatively stable performances.

Ablation on the Visual Delta Tokens. Learned Visual Delta Tokens (VDT) function as queries interacting with paired image features to extract visual difference information for LLM. To further demonstrate the role of VDT, we randomly drop a portion of them during inference to test the impact on performance of CIDEr on Birds-to-Words.

Table 2: Ablation of model design on Visual Delta Module (VDM). We conduct ablation analysis on three aforementioned benchmarks.

Layers	Tokens	Image-Editing-Request				Spot-the-Diff				Birds-to-Words			
		B	C	M	R	B	C	M	R	B	C	M	R
6	16	28.1	98.8	24.0	51.0	9.8	43.1	13.6	32.4	25.0	19.6	26.0	46.4
6	32	28.4	99.7	23.4	51.4	10.1	47.6	13.8	32.4	25.9	20.6	25.9	46.8
6	64	28.7	100.9	24.0	52.0	10.5	45.5	14.3	33.1	24.4	18.3	25.1	46.2
6	128	26.0	93.1	22.7	49.7	10.1	45.7	13.5	32.4	24.7	19.5	25.9	45.9
12	16	22.9	94.7	23.0	50.8	9.6	42.8	13.8	32.1	26.1	21.8	25.9	46.3
12	32	23.4	99.9	23.5	50.7	9.2	40.1	13.2	31.6	24.9	22.2	25.7	46.6
12	64	27.4	97.6	23.6	50.3	9.8	44.5	13.6	32.5	23.9	16.6	25.8	45.0
12	128	24.0	101.6	23.9	52.1	9.7	44.4	13.9	32.4	24.8	20.6	26.0	46.9

Table 3: Ablation of model design on Visual Delta Tokens (VDT). We conduct ablation analysis on performance of CIDEr on the Birds-to-Words benchmarks.

Used Delta Tokens	100%	50%	20%	0%
OneDiff (Vicuna-7B)	21.8	20.8	19.6	18.1

1.2 Data

This section illustrates the statistics of DiffCap and the qualitative results of data construction.

Data Statistics. Our DiffCap dataset consists of three components: collected datasets, generated data from real images and generated data from synthetic images. (See Tab. 4)

Table 4: Statistics of our DiffCap dataset. "Real" refers to our generated data from real images, and "Synthetic" refers to our generated data from synthetic images.

Component	Sub-Component	#Image-Pairs	#Annotations
Collected	Spot-the-Diff	11K	21K
	CLEVR-Change	8K	44K
	Image-Editing-Request	3.4K	4.2K
	Birds-to-Words	14K	14K
	NLVR2	86K	86K
Generated	Real	54K	54K
	Synthetic	139K	139K
Total		316K	363K

Data Samples. More samples from DiffCap are displayed as qualitative results of date construction. (See Fig. 1 and Fig. 2)

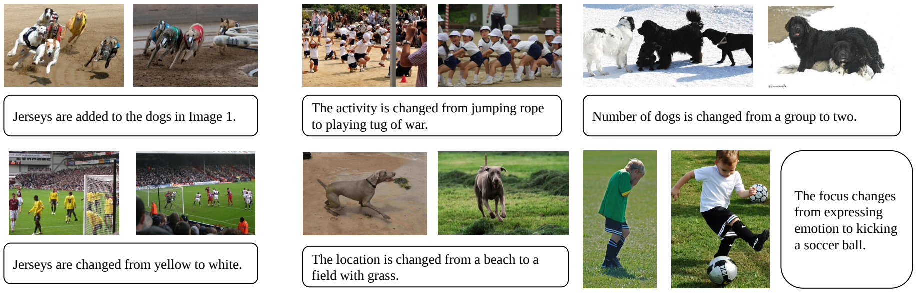

Jerseys are added to the dogs in Image 1.
The activity is changed from jumping rope to playing tug of war.
Number of dogs is changed from a group to two.
Jerseys are changed from yellow to white.
The location is changed from a beach to a field with grass.
The focus changes from expressing emotion to kicking a soccer ball.

Fig. 1: More real-image samples of our DiffCap dataset.

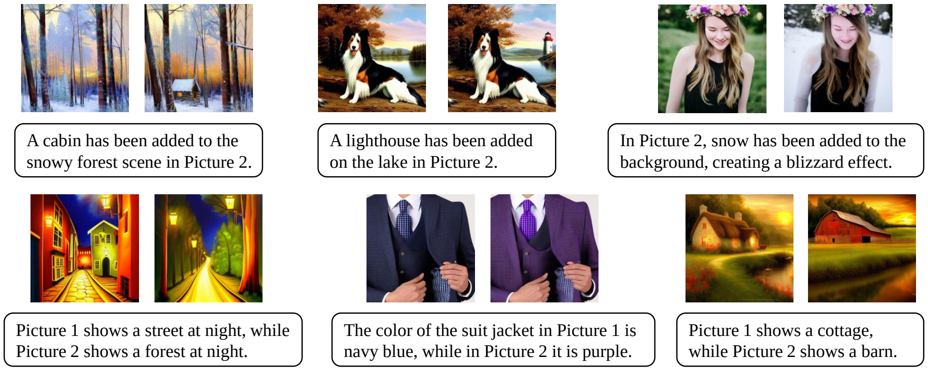
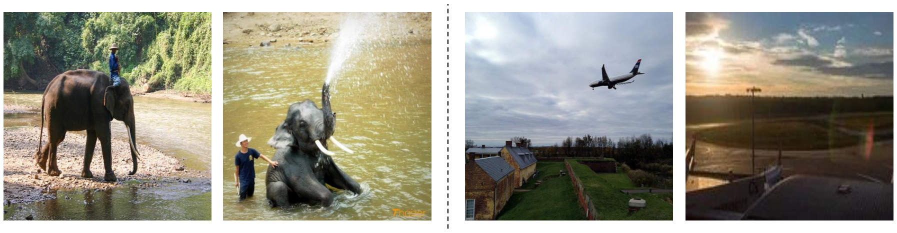

A cabin has been added to the snowy forest scene in Picture 2.
A lighthouse has been added on the lake in Picture 2.
In Picture 2, snow has been added to the background, creating a blizzard effect.
Picture 1 shows a street at night, while Picture 2 shows a forest at night.
The color of the suit jacket in Picture 1 is navy blue, while in Picture 2 it is purple.
Picture 1 shows a cottage, while Picture 2 shows a barn.

Fig. 2: More synthetic-image samples of our DiffCap dataset.

Potential Failure Cases We present potential failure cases where the environment and objectives are roughly the same, yet multiple describable image differences exist, which could lead to confusion. (See Fig. 3)



The man's position is changed from riding the elephant to standing next to it.

The main subject is changed from a jet flying over a house to a sunset over an airport.

Fig. 3: Potential failure cases of our DiffCap dataset.