Figure 1: Examples to showcase the input details when generate images using AnyText

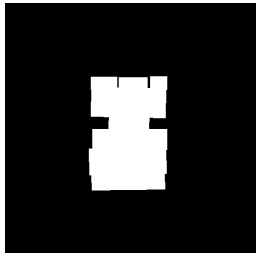
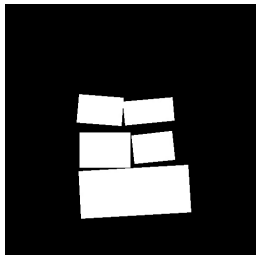
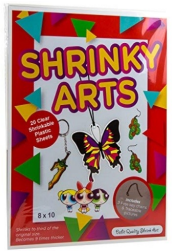
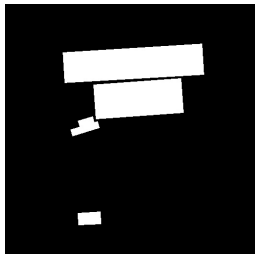
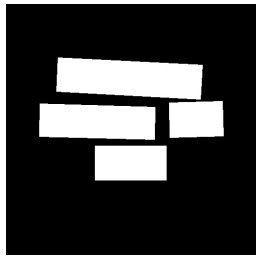
Origin	Glyph	Position	Caption
			hell hath no fury like a woman sleep deprived pullover, texts that says “DEPRIVED”, “SLEEP” , “WOMAN”, “likea” , “FURY”, “HELL”, “HATH” , “NO”.
			real men play the ukulele coffee mug, content of the text in the graphic is “ukulele”, “play” , “the”, “men” , “real”.
			owl always love you, captions shown in the snapshot are “you” , “love”, “always”, “owl” .
			a book with the title shrinky arts, captions are “8x10” , “Shrinkable” , “Clear” , “ARTS” , “SHRINKY” .
			a purple sign with the words birthday parties for kids, texts that says “Kids” , “Parties” , “for”, “Birthday”.

Figure 2: The glyph, position and caption of examples in English from the AnyText-benchmark

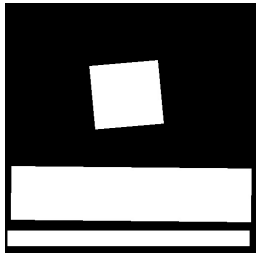
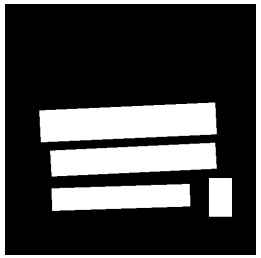
Origin	Glyph	Position	Caption
			a chinese restaurant with a sign that says jing bai hospital, texts that says “京滨招待所” .
			chinese education logo, content and position of the texts are”诚”, “诚技教育”, “CHENGJLEDUCATION” .
			a man in a suit standing at a podium in front of a screen, the written materials on the picture: “口市人民政府康美药业股份有” , “股份制医院组建启动” .
			a boy holding up a sign with chinese writing, that reads “武汉加油！ ”, “中国加油！ ”, “晟涵”, “马尾实小三年 四班” .
			wooden house number plaque with flowers, the written materials on the picture: “1086” , “支持定制” .

Figure 3: The glyph, position and caption of examples in Chinese from the AnyText-benchmark

Table 1: Quantitative comparison of AnyText and competing methods. In order to make the evaluation results more objective and fair, we replaced the OCR model from PPOCR_v3 with DuGuang_OCR. The green and red values represent the original and updated metrics, respectively.

Methods	**English**			**Chinese**		
	Sen. Acc↑	**NED↑**	**FID↓**	**Sen. Acc↑**	**NED↑**	**FID↓**
ControlNet	0.6027	0.8134	49.43	0.3742	0.6428	52.64
	0.5910	0.8061		0.3727	0.6355	
TextDiffuser	0.5879	0.7974	43.91	0.0596	0.1233	52.55
	0.5878	0.7958		0.0600	0.1248	
GlyphControl	0.3379	0.6488	36.49	0.0324	0.0807	32.04
	0.3382	0.6390		0.0301	0.0775	
GlyphControl (TextCaps-5k)	0.4906	0.7392	43.35	0.0406	0.0915	48.17
	0.4933	0.7324		0.0412	0.0939	
AnyText	0.6686	0.8634	**35.71**	0.7090	0.8877	**28.69**
	0.6509	**0.8541**		**0.6646**	**0.8274**	

Table 2: Ablation experiments of AnyText on a small-scale dataset from AnyWord-3M. We have added two additional experiments: Exp. 3 and Exp. 4. In Exp. 3, we replaced the OCR model in the text embedding module with CLIP vision model (vit), and in Exp. 4, we replaced it with a stacked convolutional module (conv). Additionally, all ablation experiments' Sen. Acc and NED metrics were recalculated using the DuGuang_OCR model.

Exp. №	Editing	Position	Text Embedding	Perceptual Loss	λ	Chinese	
						Sen. Acc↑	NED↑
1	✓	✓	✗	✗	-	0.1552	0.4070
2	✗	✓	✗	✗	-	0.2024	0.4649
3	✗	✓	vit	✗	-	0.1416	0.3809
4	✗	✓	conv	✗	-	0.1864	0.4402
5	✗	✓	ocr	✗	-	0.4595	0.7072
6	✗	✗	ocr	✗	-	0.4472	0.6974
7	✗	✓	ocr	✓	0.003	0.4848	0.7353
8	✗	✓	ocr	✓	0.01	0.4996	0.7457
9	✗	✓	ocr	✓	0.03	0.4659	0.7286

Figure 4: Qualitative comparison of ablation experiments from the AnyText-benchmark in English

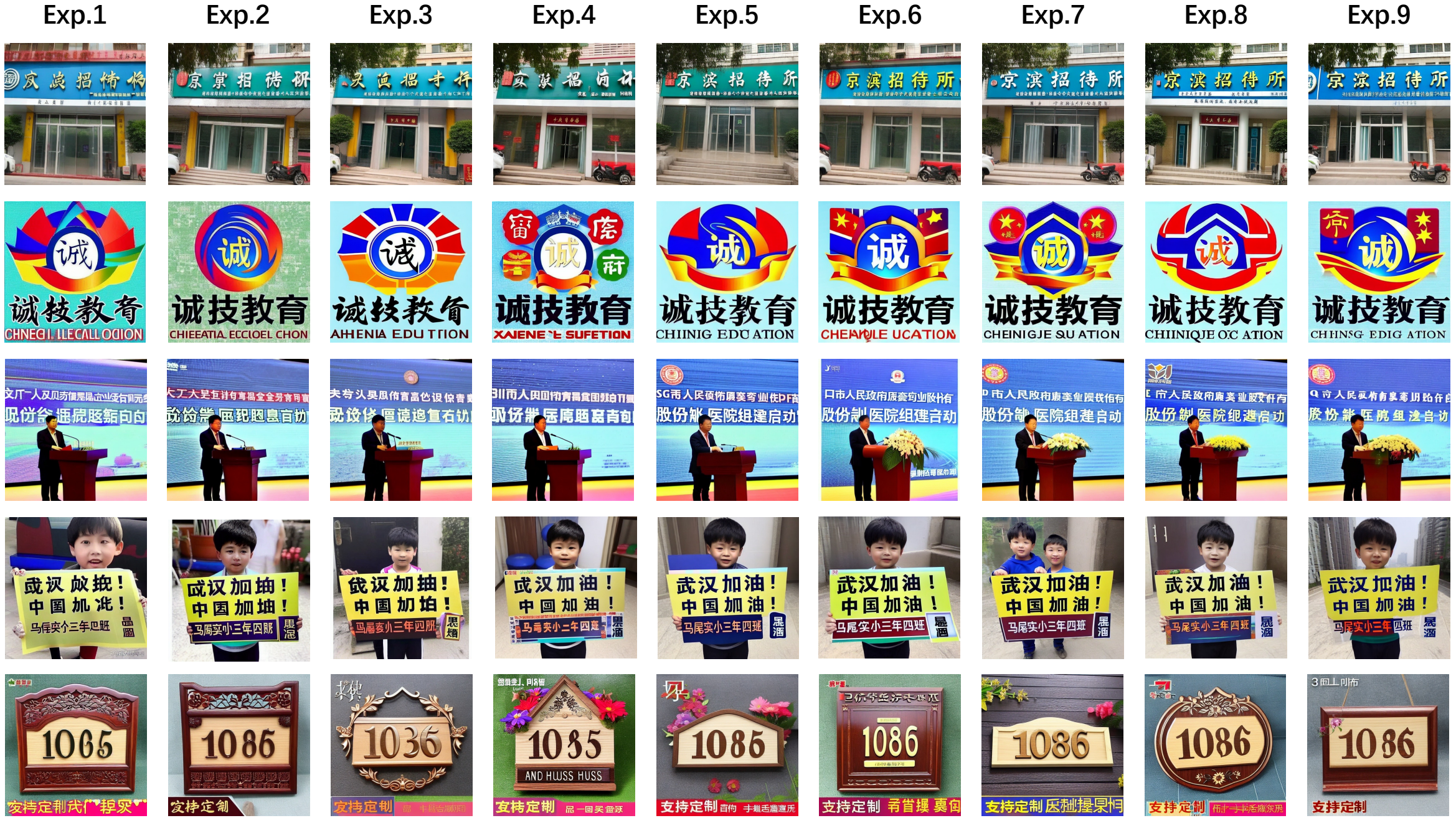


Figure 5: Qualitative comparison of ablation experiments from the AnyText-benchmark in Chinese

t	φ (t)	x'_0
900	0.0142	
700	0.0825	
500	0.2763	
300	0.5905	
100	0.8942	

Figure 6: Visualization of x'_0 at different time steps **t**, along with the values of the weight adjustment function φ**(t)**.

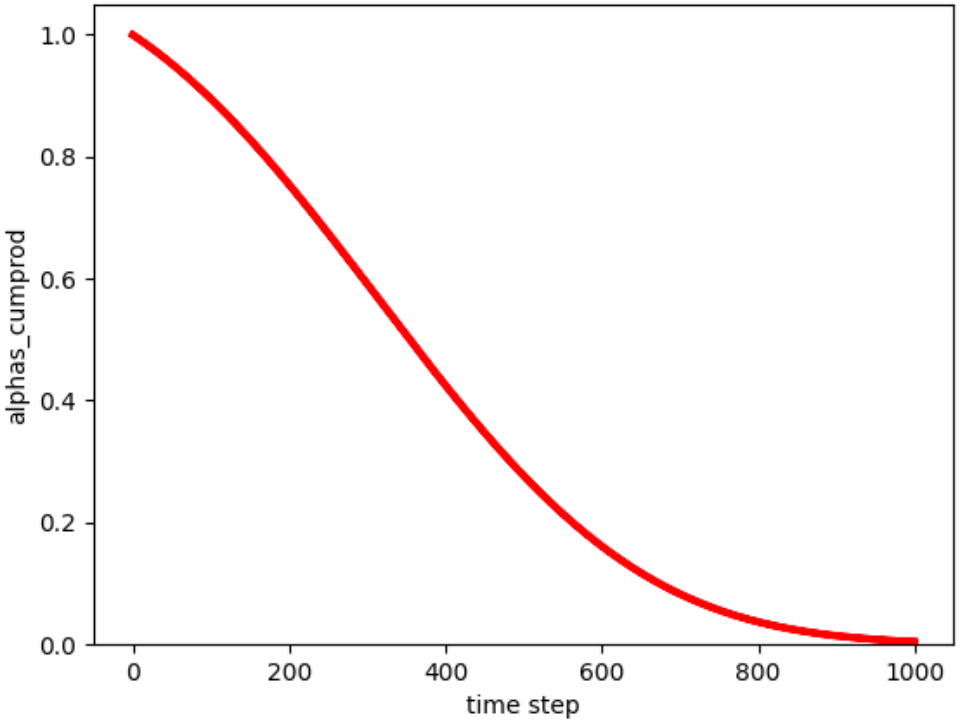


Figure 7: Curve depicting the changing of $\bar{\alpha}_t$ with time step t.

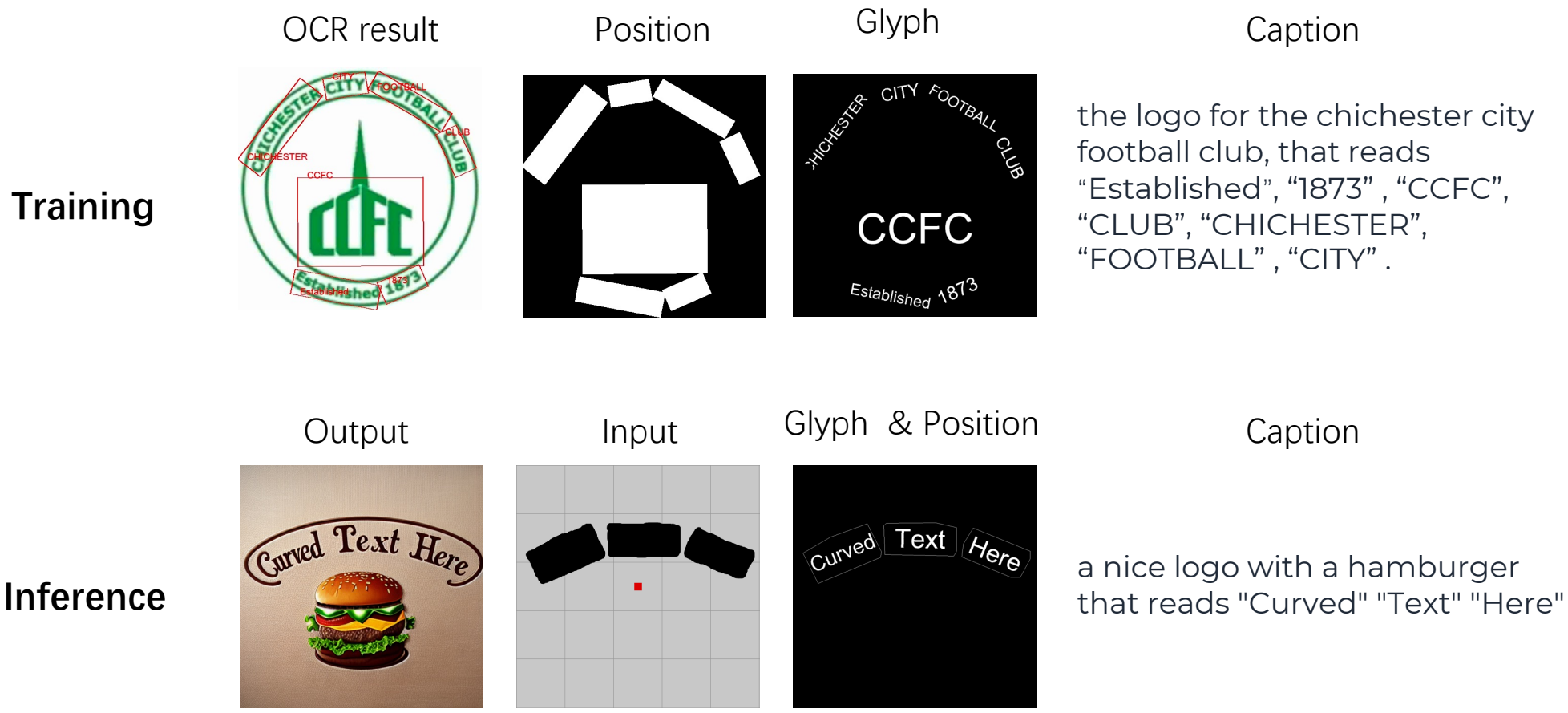


Figure 8: Illustration demonstrating how the model can learn the correspondence between curved rectangle bounding boxes and curved text.