A. Additional Results

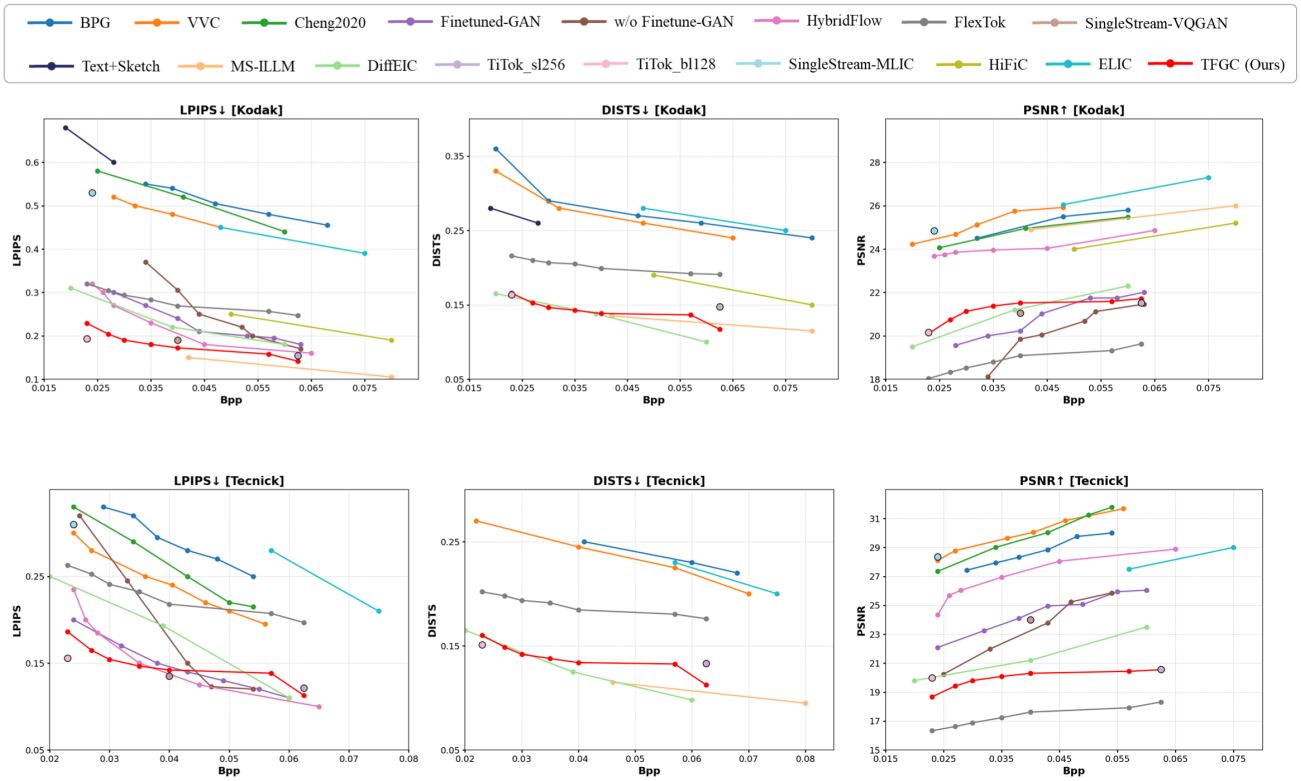


Figure 7. Quantitative comparison using PSNR↑, LPIPS↓ [73], DISTS↓ [13] on the Kodak [18] and Tecnick [2] datasets. Results for BPG [5], VVC [7], Cheng2020 [11], Finetuned-GAN [50], w/o Fine-tune [50], HybridFlow [47], FlexTok [3], ELIC [23], HiFiC [52], Text+Sketch [33], MS-ILLM [55], DiffEIC [41], TiTok_sl256 [69], TiTok_bl128 [69], SingleStream-MLIC [28], SingleStream-VQGAN [15]. Baseline numbers were taken from the [41, 47].

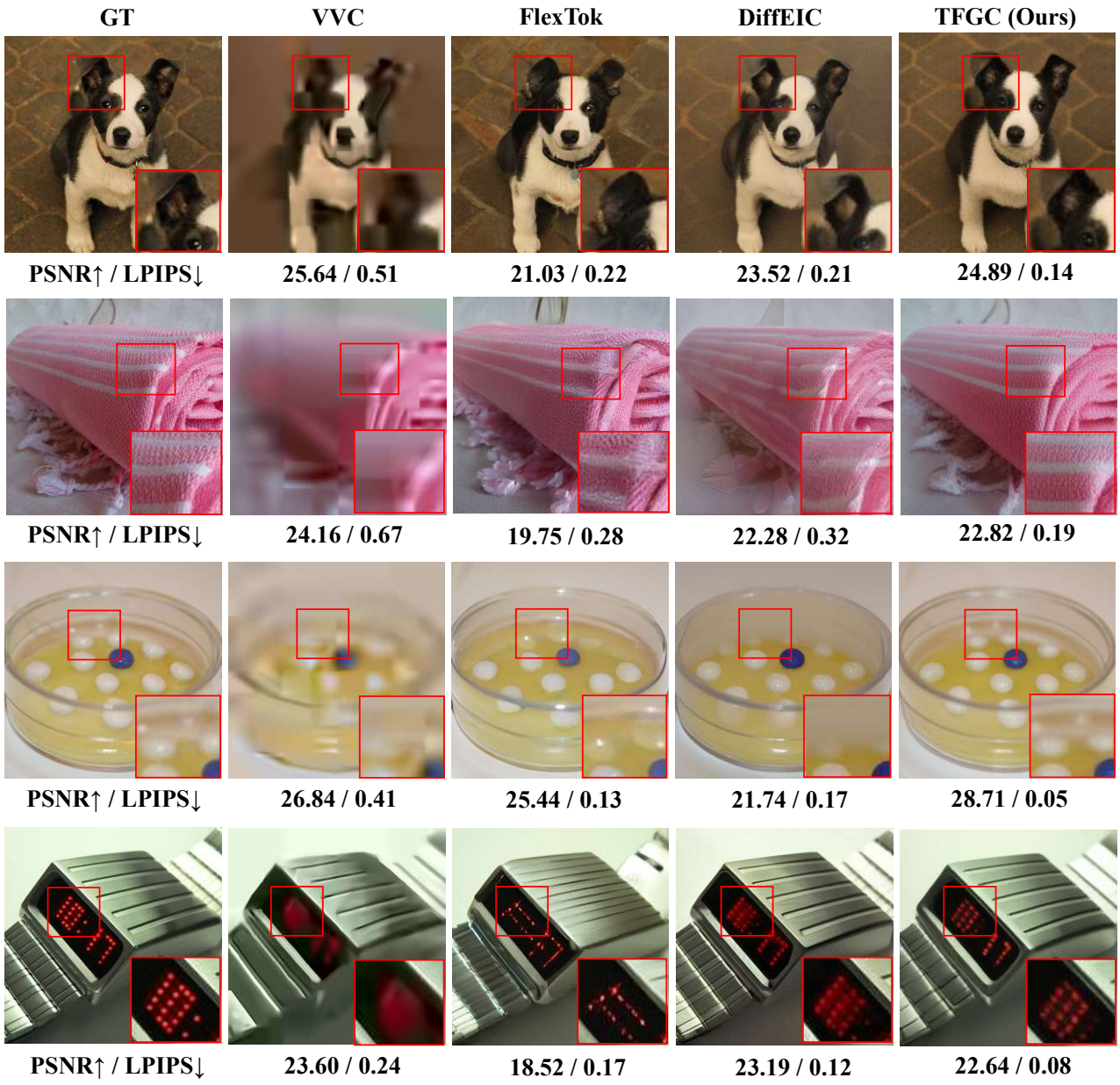


Figure 8. Visualization results of our method and other baselines at 0.04 bpp on human-oriented benchmarks. GT denotes the ground-truth. The corresponding PSNR↑ and LPIPS↓ metrics are reported above.

Table 7. Results on Caption benchmarks. "Var." denotes whether a single model supports variable bitrate control, and "Ave. Bpp" represents the average bits per pixel across all datasets. Each group corresponds to a specific bitrate range, with bitrate variation constrained within $\pm$0.005 bpp. **Bold** and underlined denote the best and second-best results, respectively.

Methods	Ave. Bpp	Var.	Param.	MSCOCO			Flickr30k		
				BLEU-4↑	ROUGE-L↑	CIDEr↑	BLEU-4↑	ROUGE-L↑	CIDEr↑
VVC [7]	0.065	–	–	7.36	33.44	14.31	7.51	32.02	12.33
TiTok$_{sl256}$ [69]	0.063	✗	330M	13.45	42.31	33.53	11.10	37.06	17.15
DiffEIC [41]	0.062	✗	1380M	11.42	39.00	16.45	12.45	38.60	19.75
ELIC [23]	0.059	✗	34M	11.80	39.80	23.22	–	–	–
FlexTok [3]	0.063	✓	950M	10.91	39.10	22.27	10.52	36.78	20.55
TFGC (Ours)	0.063	✓	332M	**24.90**	**49.94**	**82.73**	**14.12**	**39.04**	**27.29**
VVC [7]	0.045	–	–	5.63	31.34	14.31	6.55	30.34	10.56
DiffEIC [41]	0.040	✗	1380M	11.03	38.60	15.63	11.53	37.25	17.83
FlexTok [3]	0.035	✓	950M	10.60	38.62	21.51	10.11	36.25	18.68
TFGC (Ours)	0.035	✓	332M	**24.60**	**49.76**	**81.81**	**13.59**	**38.72**	**26.97**
VVC [7]	0.024	–	–	2.40	26.40	3.12	2.83	25.51	3.34
TiTok$_{bl128}$ [69]	0.023	✗	390M	12.68	41.43	31.99	10.23	35.93	15.88
DiffEIC [41]	0.022	✗	1380M	10.52	37.90	14.52	9.66	34.44	13.85
FlexTok [3]	0.023	✓	950M	10.22	38.07	21.08	9.89	35.63	17.52
TFGC (Ours)	0.023	✓	332M	**23.09**	**48.32**	**74.96**	**12.80**	**37.60**	**24.00**

Table 8. Results on Vision Grounding benchmarks. "Var." denotes whether a single model supports variable bitrate control, and "Ave. Bpp" represents the average bits per pixel across all datasets. Each group corresponds to a specific bitrate range, with bitrate variation constrained within $\pm$0.005 bpp. **Bold** and underlined denote the best and second-best results, respectively.

Methods	Ave. Bpp	Var.	Param.	Refcoco (Acc@0.5↑)		Refcocog (Acc@0.5↑)	
				val	test	val	test
VVC [7]	0.063	–	–	26.53	24.83	19.30	19.06
TiTok$_{sl256}$ [69]	0.063	✗	330M	51.24	50.72	44.57	43.88
DiffEIC [41]	0.063	✗	1380M	**62.19**	**61.93**	**54.49**	54.09
ELIC [23]	0.063	✗	34M	56.23	54.89	47.39	47.00
FlexTok [3]	0.063	✓	950M	57.00	53.82	45.77	46.16
TFGC (Ours)	0.063	✓	332M	61.49	61.27	52.29	**54.35**
VVC [7]	0.045	–	–	21.41	19.79	14.77	14.47
DiffEIC [41]	0.040	✗	1380M	59.80	59.31	50.71	51.06
FlexTok [3]	0.035	✓	950M	55.15	51.46	43.87	43.56
TFGC (Ours)	0.035	✓	332M	**61.22**	**60.92**	**52.59**	**54.62**
VVC [7]	0.023	–	–	9.78	9.91	7.21	6.06
TiTok$_{bl128}$ [69]	0.023	✗	390M	48.86	49.08	41.73	42.06
DiffEIC [41]	0.021	✗	1380M	54.81	**55.18**	46.20	45.09
FlexTok [3]	0.023	✓	950M	54.22	49.29	42.67	41.88
TFGC (Ours)	0.023	✓	332M	**54.96**	54.89	**47.22**	**47.51**

Table 9. Results on Hallucination benchmarks. "Var." denotes whether a single model supports variable bitrate control, and "Ave. Bpp" represents the average bits per pixel across all datasets. Each group corresponds to a specific bitrate range, with bitrate variation constrained within ±0.005 bpp. **Bold** and underlined denote the best and second-best results, respectively.

Methods	Ave. Bpp	Var.	Param.	POPE		
				F1-score↑	Accuracy↑	Precision↑
VVC [7]	0.065	–	–	70.22	72.33	77.43
TiTok$_{sl256}$ [69]	0.063	✗	330M	78.28	80.45	89.44
DiffEIC [41]	0.061	✗	1380M	80.64	82.05	88.90
ELIC [23]	0.060	✗	34M	79.78	81.38	88.63
FlexTok [3]	0.063	✓	950M	77.11	79.06	86.20
TFGC (Ours)	0.063	✓	332M	**85.96**	**86.37**	**89.61**
VVC [7]	0.045	–	–	66.44	68.46	72.24
DiffEIC [41]	0.039	✗	1380M	79.33	81.20	89.44
FlexTok [3]	0.035	✓	950M	74.51	77.17	85.65
TFGC (Ours)	0.035	✓	332M	**85.94**	**86.34**	**89.50**
VVC [7]	0.025	–	–	61.62	63.17	65.41
TiTok$_{bl128}$ [69]	0.023	✗	390M	77.34	79.53	**88.30**
DiffEIC [41]	0.021	✗	1380M	75.24	77.95	84.16
FlexTok [3]	0.023	✓	950M	74.92	77.38	85.38
TFGC (Ours)	0.023	✓	332M	**82.57**	**83.12**	86.34

Table 10. **Results on machine-oriented benchmarks.** We also evaluate machine understanding using **InternVL3-1B** [10] as the off-the-shelf visual-language model in 0.02 bpp. "Var." denotes whether a single model supports variable bitrate control, and "Ave. Bpp" represents the average bits per pixel across all datasets. Caption performance is evaluated on the MSCOCO, VQA performance on QKVQA, and Grounding performance using the Acc@0.5 metric. For all metrics, higher is better. **Bold** and underlined denote the best and second-best results, respectively.

Methods	Ave. Bpp	Var.	Param.	Caption		Grounding		VQA	
				ROUGE-L	CIDEr	RefCOCO	RefCOCOg	Acc	BLEU-4
VVC [7]	0.025	–	–	33.57	12.20	18.50	13.96	19.24	4.94
TiTok$_{bl128}$ [69]	0.023	✗	390M	53.84	99.21	62.82	57.10	32.45	7.79
DiffEIC [41]	0.021	✗	1380M	52.73	96.15	64.51	52.79	37.78	9.04
FlexTok [3]	0.023	✓	950M	51.17	84.26	57.30	51.92	29.73	7.22
TFGC (Ours)	0.023	✓	332M	**54.19**	**99.62**	**80.36**	**63.40**	**39.43**	**9.53**

Prompt: Provide a one-sentence caption for the provided image.

VVC: A blurry view of a kitchen counter.

TiTok: A white computer mouse sits on a desk next to a beige white keyboard.

FlexTok: A silver and white computer mouse sits on a desk next to a keyboard.

ELIC: A computer with a silver-colored keyboard and a white mouse.

DiffEIC: A keyboard with a gray color scheme, placed on a desk with a mouse nearby it.

TFGC: A white keyboard on a desk. A mouse is on top of the keyboard.

Figure 9. Visualization examples of our proposed method and other baselines in the 0.06 bpp range for the image Captioning task. Model configurations are consistent with those in Table 7. Red text indicates incorrect content, green denotes correct content, and orange highlights ambiguous or unclear expressions.

Prompt: Provide a one-sentence caption for the provided image.

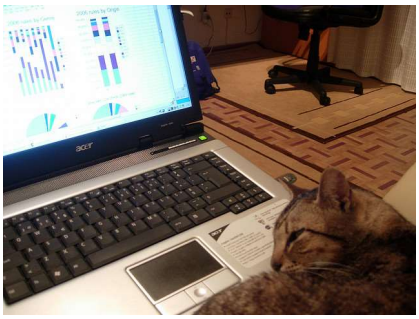
VVC: A blurry shot of a room with a white light on the wall.

FlexTok: woman is standing in front of a window, looking out and adjusting the curtain.

DiffEIC: Two women are standing in a living room, facing a television screen.

TFGC: A woman is playing a video game on a computer, while the other stands beside her.

Figure 10. Visualization examples of our proposed method and other baselines in the 0.04 bpp range for the image Captioning task. Model configurations are consistent with those in Table 7. Red text indicates incorrect content, green denotes correct content, and orange highlights ambiguous or unclear expressions.



Prompt: Provide a one-sentence caption for the provided image.

VVC: A man is sitting in front of a computer.

TiTok: A cat is lying on a beige rug in front of a laptop computer.

FlexTok: A cat is sitting on a laptop keyboard, looking at the screen.

DiffEIC: A person is sitting on a bed, working on a laptop with a tablet in front of them.

TFGC: A cat is laying on a keyboard and mouse.

Figure 11. Visualization examples of our proposed method and other baselines in the 0.02 bpp range for the image Captioning task. Model configurations are consistent with those in Table 7. Red text indicates incorrect content, green denotes correct content, and orange highlights ambiguous or unclear expressions.

Prompt: Please provide the bounding box coordinate of the region this sentence describes: bowl on right.

VVC: 0.1668 DiffEIC: 0.6610 FlexTok: 0.7922 TFGC: 0.8554

Figure 12. Visualization examples of our proposed method and other baselines in the 0.04 bpp range for the image Grounding task. Model configurations are consistent with those in Table 7. The red box denotes the GT bounding box, the blue box represents the predicted box, and the IoU [17] values are reported in the figure.

Figure 13. Visualization examples of our proposed method and other baselines in the 0.02 bpp range for the image Grounding task. Model configurations are consistent with those in Table 7. The red box denotes the GT bounding box, the blue box represents the predicted box, and the IoU [17] values are reported in the figure.

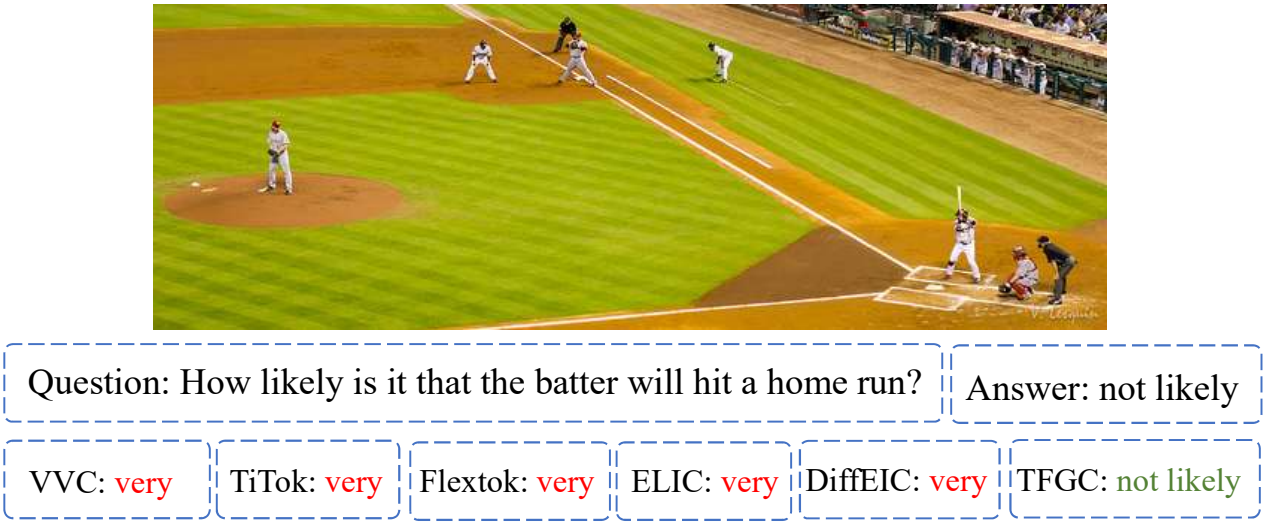


Figure 14. Visualization examples of our proposed method and other baselines in the 0.06 bpp range for the image VQA task. Model configurations are consistent with those in Table 7.

Question: What room of the house is this?

Answer: living room

VVC: kitchen

Flextok: dining room

DiffEIC: dining room

TFGC: living room

Figure 15. Visualization examples of our proposed method and other baselines in the 0.04 bpp range for the image VQA task. Model configurations are consistent with those in Table 7.

Question: Is there a ball in this photo?

Answer: no

VVC: yes

TiTok: yes

Flextok: yes

DiffEIC: yes

TFGC: no

Figure 16. Visualization examples of our proposed method and other baselines in the 0.02 bpp range for the image VQA task. Model configurations are consistent with those in Table 7.

Question: Is there an orange in the imange?

Answer: no

VVC: yes | TiTok: yes | Flextok: yes | ELIC: yes | DiffEIC: yes | TFGC: no

Figure 17. Visualization examples of our proposed method and other baselines in the 0.06 bpp range for the Hallucination effects. Model configurations are consistent with those in Table 7.

Question: Is there a knife in the image?

Answer: yes

VVC: no | Flextok: no | DiffEIC: no | TFGC: yes

Figure 18. Visualization examples of our proposed method and other baselines in the 0.04 bpp range for the Hallucination effects. Model configurations are consistent with those in Table 7.

Question: Is there a couch in the image?

Answer: no

VVC: yes

TiTok: yes

Flextok: yes

DiffEIC: yes

TFGC: no

Figure 19. Visualization examples of our proposed method and other baselines in the 0.02 bpp range for the Hallucination effects. Model configurations are consistent with those in Table 7.