

Low-Bitrate Video Compression through Semantic-Conditioned Diffusion

Supplementary Material

A. Video Visualization

We provide additional video samples to facilitate a more comprehensive comparison among codecs, extending the analysis presented in Fig. 5. Specifically, we showcase the ground truth video alongside compression results from all reproducible approaches, including our proposed DiSCo method, DCVC-RT [26], DCVC-FM [32], H.266 [3], H.265, and H.264. The visualizations reveal that at low bitrates, such as 0.005 BPP, DCVC-RT and DCVC-FM often suffer from blurry artifacts and a lack of detail due to the aggressive quantization of intermediate features. Meanwhile, H.266 typically exhibits blocky and motion-related artifacts caused by its block-based encoding and motion prediction mechanisms. In contrast, our method delivers significantly superior perceptual quality and temporal consistency, preserving both high semantic fidelity and pixel-level accuracy relative to the ground truth.

Additionally, we present video samples of our method operating at ultra-low bitrates in extension to Fig. 11. Even when the degraded reference video becomes barely recognizable, our approach consistently generates content with high visual quality and acceptable semantic accuracy. This demonstrates the effectiveness of our complementary semantic modalities and spatiotemporal generation.

Finally, we provide a video demonstration showcasing the impact of token interleaving. We include the interleaved reference video in pixel space, the reconstructed video conditioned on the multimodal reference, and the ground truth video. The results illustrate that for both sketch and pose modalities, our method is capable of reconstructing content with exceptional quality and temporal coherence.

B. Multimodal Interleaving

We found that the token level is the minimum appropriate granularity for the multimodal interleaving strategy. At a finer granularity of the frame level, we can interleave 1 RGB video frame followed by several auxiliary modality frames. However, this strategy leads to an undesirable mixture of information within the latent space. We prove this by reconstructing a video that interleaves 1 RGB frame and 7 pose frames. In Fig. 12, we present the sampled RGB video frames before VAE encoding on the left, and the same frames after VAE decoding on the right. Visualization shows that the pose skeleton leaks into the RGB frame within the latent space, severely compromising its fidelity. Consequently, it leads to substantial artifacts in the conditional generated result.

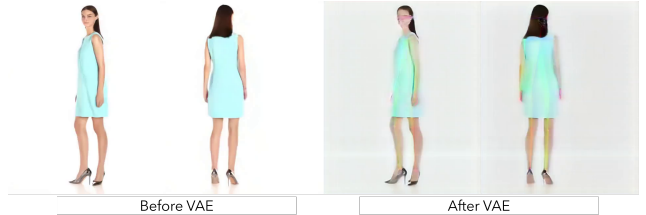


Figure 12. Modality mixture caused by frame interleaving.

Table 8. Bitrate allocation across modalities under a total bandwidth of 50 kbps ($QP=0$, $D_s=1$, $D_t=2$). Text bitrate is used as the reference unit.

Modality	Bitrate (bps)	Ratio
Text	1249.5	1.00
Pose	3111.0	2.49
Sketch	16216.2	12.98
Degraded Video	32015.7	25.62

Table 9. Per-frame latency breakdown (ms/frame).

Component	Extract	Encode	Decode
Video	N/A	6.98	5.68
Text (amortized)	81.40	0.01	0.02
Sketch	9.51	6.98	5.68
Pose	11.70	0.30	0.01
Diffusion	390.52 (13B) / 85.08 (2B)		

C. Multimodal Bitrate Allocation

Given $QP=0$, $D_s=1$, $D_t=2$, 512×512 resolution, 57 frames, and 30 fps, the total bandwidth is approximately 50 kbps. The allocation of bitrate across all modalities, along with their relative ratios compared to the text modality, is presented in Tab. 8. Note that the cost of sending the text description is a one-time expense. We amortize it across each frame to calculate the bitrate. The measurement reveals that the degraded video consumes the majority of the overall bitrate, while the auxiliary modalities are significantly more compact. This outcome aligns with our design intuition to leverage the compactness of multimodal semantics and demonstrates the effectiveness of our proposed modality-specific codecs.

D. Efficiency Analysis

On an H100 GPU, finetuning a 13B-parameter Diffusion decoder requires approximately 2 hours for latent preprocessing and 2.5 hours for training, consuming 49 GB of GPU memory. During inference, the total encoding latency is about 97.89 ms per frame, while the decoding latency is 396.20 ms. A detailed latency breakdown of our method is available at Tab. 9. The Diffusion decoder can also be finetuned over a 2B distilled backbone. It will further reduce the training time to 1 hour and the decoding latency to 85.08 ms, at the cost of degrading video quality by 5.45% in PSNR, 11.86% in LPIPS, and 35.97% in FVMD.

E. Training Details

Our 13B-parameter Diffusion decoder is finetuned from LTX-Video backbone `LTXV_13B_098_DEV` using LoRA adaptation, while the 2B version is finetuned from `LTXV_2B_0.9.8_DISTILLED`. Specifically, LoRA is applied with rank 256 and scaling factor 256 to the attention projection layers (`to_q`, `to_k`, `to_v`, `to_out`) and the feed-forward layers (`ff.net.0.proj`, `ff.net.2`). The training is performed on videos with spatial resolution 512×512 and 57 frames.

We optimize the model for 8,000 steps using the AdamW optimizer with a learning rate of 2×10^{-4} and a cosine learning-rate schedule. The batch size is 1, and gradient checkpointing is enabled to reduce memory usage. All training runs use bfloat16 mixed-precision. Flow-matching training follows a shifted logit-normal timestep sampling strategy. The model is trained on the 8000-video subset from OpenVid-1M dataset using preprocessed latent representations to accelerate training. During inference, we use classifier-free guidance with scale 3.5 and 50 diffusion steps to generate videos.

F. Quantitative Performance

For an easy quantitative comparison with our method, we provide the detailed performance under various test settings. These results are presented in Tab. 10 for the HEVC-B dataset, Tab. 11 for the MCL-JCV dataset, and Tab. 12 for the UVG dataset, respectively. All videos are resized to 512×512 -resolution 57-frame clips during evaluation.

Table 10. Rate-distortion performance of DiSCo on the HEVC-B dataset.

QP	D_s	D_t	BPP	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	FID \downarrow	FVD \downarrow	FVMD \downarrow	FloLPIPS \downarrow
0	4	8	0.0005	17.33	0.4057	0.5104	0.2140	231.28	2166.10	4176.84	0.5395
0	2	4	0.0010	20.18	0.5141	0.3789	0.1744	169.36	985.70	1485.72	0.4358
0	1	8	0.0019	21.50	0.5911	0.3158	0.1707	150.88	748.50	1489.55	0.3714
0	1	4	0.0024	22.27	0.6036	0.2782	0.1500	132.17	586.30	1057.73	0.3384
0	1	2	0.0031	23.15	0.6358	0.2571	0.1468	123.55	489.30	834.30	0.3157
8	1	2	0.0049	24.44	0.6901	0.2114	0.1344	106.36	396.30	705.02	0.2774
16	1	2	0.0078	25.54	0.7364	0.1735	0.1266	92.63	291.90	620.90	0.2451
24	1	2	0.0121	26.45	0.7705	0.1438	0.1178	80.68	233.20	540.62	0.2150
32	1	2	0.0190	27.05	0.7959	0.1230	0.1102	72.73	204.90	490.08	0.1913

Table 11. Rate-distortion performance of DiSCo on the MCL-JCV dataset.

QP	D_s	D_t	BPP	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	FID \downarrow	FVD \downarrow	FVMD \downarrow	FloLPIPS \downarrow
0	4	8	0.0002	18.52	0.4875	0.4717	0.2338	248.55	1898.10	9082.96	0.4727
0	2	4	0.0007	21.60	0.5957	0.3647	0.1998	198.99	1072.90	5460.21	0.3779
0	1	8	0.0016	22.49	0.6371	0.3108	0.1831	169.57	894.90	4817.98	0.3304
0	1	4	0.0021	23.49	0.6555	0.2768	0.1669	154.98	679.10	4398.26	0.3003
0	1	2	0.0029	24.60	0.6855	0.2565	0.1624	149.12	593.40	2183.66	0.2812
8	1	2	0.0046	26.05	0.7357	0.2155	0.1513	130.96	483.10	1909.99	0.2447
16	1	2	0.0074	27.07	0.7702	0.1842	0.1414	116.70	400	1662.79	0.2163
24	1	2	0.0116	28.01	0.7978	0.1575	0.1307	101.03	336.20	1512.88	0.1923
32	1	2	0.0182	28.71	0.8184	0.1414	0.1249	94.34	262.20	1311.55	0.1725

Table 12. Rate-distortion performance of DiSCo on the UVG dataset.

QP	D_s	D_t	BPP	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	DISTS \downarrow	FID \downarrow	FVD \downarrow	FVMD \downarrow	FloLPIPS \downarrow
0	4	8	0.0002	18.26	0.5008	0.4514	0.2295	193.09	1673.30	15403.81	0.4780
0	2	4	0.0008	20.72	0.5831	0.3608	0.1949	148.74	1144.10	8494.23	0.3992
0	1	8	0.0017	21.74	0.6354	0.3121	0.1779	123.80	853.20	8862.03	0.3498
0	1	4	0.0023	22.39	0.6447	0.2870	0.1641	117.34	683	6837.15	0.3257
0	1	2	0.0031	23.43	0.6662	0.2688	0.1583	115.85	651.30	5398.95	0.3133
8	1	2	0.0050	24.63	0.7115	0.2309	0.1463	99.76	527.20	5040.25	0.2860
16	1	2	0.0080	25.53	0.7424	0.2013	0.1360	89.11	448.60	4588.12	0.2559
24	1	2	0.0126	26.35	0.7676	0.1798	0.1284	82.49	391.10	4287.29	0.2338
32	1	2	0.0196	26.94	0.7875	0.1677	0.1246	79.37	371.60	4086.40	0.2153