SweetTok: Semantic-Aware Spatial-Temporal Tokenizer for Compact Video Discretization

Supplementary Material

Notations	Explanations
$DQAE_{s,t}$	Decoupled query autoencoder
$\mathcal{P}$	Patchify Module
$\mathcal{E}$	Encoder
$\mathcal{D}$	Decoder
$\mathcal{Q}$	Quantizer
$p_{t,h,w}$	Downsample ratio
$\mathbf{Q}_{s,t,m}$	Continuous latent query tokens
$\mathbf{Z}_{\mathbf{Q}_{s,t,m}}$	Continuous latent query token
$\tilde{\mathbf{Q}}_{s,t}$	Quantized latent query tokens
$\tilde{\mathbf{Z}}_{\mathbf{Q}_{s,t,m}}$	Quantized latent query token
$v_{s,t}$	Continuous visual feature
$\tilde{v}_{s,t}$	Discrete visual feature
Δv	Difference of consecutive features
$C_{adj,noun,adverb,verb}$	Codebook text embeddings
$c_{adj,noun,adverb,verb}$	Codebook text embedding
$\mathcal{F}$	Projector network
$sg(\cdot)$	Stop gradient operation

Table 1. Explanations for the notations in the main paper.

1. Experimental Settings

1.1. Model Implementation Details

Visual Tokenizer. The tokenizer is composed of an encoder $\mathcal{E}$, decoder $\mathcal{D}$, and latent quantizer $\mathcal{Q}$. The tokenizer takes a video clip of 17 consecutive frames with a resolution of 256×256 with the elements normalized to $[-0.5, 0.5]$ as input. Then the video clip will be patchified to a resolution of $1 \times 32 \times 32$ spatial feature and $4 \times 32 \times 32$ temporal feature as illustrated in the main paper. The encoder $\mathcal{E}$ and decoder $\mathcal{D}$ in our tokenizer are both composed of 8 $DQAE_s$ modules and 4 $DQAE_t$ modules with 512 hidden states and 8 attention heads. Each modules consists of self-attention, feed-forward and cross-attention layers. Before the attention computation, the visual features will be reshaped into $[(BT) \times (HW) \times D]$ and $[(BHW) \times T \times D]$ for $DQAE_s$ and $DQAE_t$ modules, respectively. The encoder $\mathcal{E}$ generates 256 spatial and 1024 temporal continuous latent tokens. These tokens are then passed to the quantizer $\mathcal{Q}$, which produces the quantized spatial and temporal latent tokens. The quantizer $\mathcal{Q}$ is composed of a spatial and temporal codebook and a GCN with two hidden layers with hidden dimension of 512 as the projector network $\mathcal{F}$. To improve the training stability of the visual tokenizer, we adopt exponential moving average (EMA) updates with weight of 0.999 following [5].

Tokenizer	PSNR ↑	SSIM ↑	LPIPS ↓
UCF-101			
LARP [2]	27.88	-	0.085
SweetTok	**29.27**	**0.7766**	**0.070**
ImageNet			
LlamaGen-16 [4]	20.79	0.675	-
TokenFlow [3]	21.41	0.687	-
LlamaGen-8 [4]	24.45	0.813	-
SweetTok	**30.23**	**0.826**	**0.068**

Table 2. More evaluation results on UCF-101 and ImageNet.

Language Model. We utilize VideoGPT following [1] as the default large language model for the video generative pre-training. All settings follow the protocol of [1].

1.2. Training Datasets

UCF-101. UCF-101 is a large-scale action recognition dataset consisting of 13,320 videos with 9537 for training and 3783 for testing across 101 action categories. The dataset includes videos with significant variations in camera motion, object appearance, scale, viewpoint, cluttered backgrounds, and lighting conditions, making it one of the most challenging datasets for action recognition.

Kinetic-600. Kinetics-600 is a large-scale action recognition dataset containing approximately 480K videos across 600 action categories. The dataset is split into 390K training, 30K validation, and 60K test videos. Each video is a 10-second clip extracted from raw YouTube footage, focusing on key action moments.

ImageNet-1K. ImageNet-1K is a widely used subset of the larger ImageNet dataset, specifically designed for image classification tasks. It contains 1.2 million labeled images across 1,000 distinct categories, ranging from animals and plants to everyday objects and scenes. Each category in ImageNet-1K includes a set of training images, along with separate validation and test sets for model evaluation. The dataset is widely used for researchs in computer vision.

1.3. Notations

The meaning of our notations appeared in the main paper are explained in Table 1.

Configuration	Language Model	Tokenizer	
		Image Finetune	Video Training
LLM init	VideoGPT	-	-
Optimizer	AdamW	AdamW	AdamW
Optimizer Hyperparameters	$\beta_1 = 0.9$, $\beta_2 = 0.96$	$\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 1e^{-8}$	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Batch size per GPU	4	32	12
Peak learning rate	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$
Discriminator peak learning rate	-	-	$1e^{-4}$
Learning rate schedule	Cosine	Cosine	Cosine
Training steps	1000K	500K	1000K
Discriminator start steps	-	-	20K
Warm-up steps	10K	10K	10K
Weight decay	0.03	$1e^{-4}$	$1e^{-4}$
Numerical precision	float16	float16	bfloat16

Table 3. The detailed training hyperparameters of SweetTok.

1.4. Training Settings

The detailed training hyper-parameter settings for SweetTok are reported in Table 3. During video training, we train the first 500K steps using proxy code following [5] to accelerate training. The remaining 500K is trained without proxy code.

2. Additional Results

2.1. More evaluation metrics

We assess SweetTok using additional metrics: PSNR, SSIM, and LPIPS. As shown in Table 2, SweetTok outperforms all baselines on both video and image datasets, further validating the superiority of our model design.

2.2. More Visualizations

Fig 3 and Fig 2 visualize the reconstruction results for the UCF-101 and K-600 datasets. The pixel-level differences between ground truth and model are shown, with brighter areas indicating greater disparity and darker areas reflecting consistency. As shown, SweetTok exhibits fewer reconstruction differences compared to OmniTokenizer, demonstrating its superior performance.

Fig 3 visualize the reconstruction results of SweetTok on ImageNet-1K. For reconstruction, differences between models are highlighted in red blocks, with details shown in green blocks. Clearly, SweetTok outperforms all baselines by a significant margin.

Finally, we visualize the words from our MLC in Fig 4, based on few-shot video action recognition tasks on the UCF-101 dataset. We use adjectives, nouns, adverbs, and verbs as prompts to Qwen LLM for action prediction. Green and orange indicate meaningful words, while red marks meaningless ones. The visualization shows that correct verb words consistently lead to accurate predictions, even when other words are irrelevant, highlighting the importance of our MLC modules for video action recognition.

3. Limitations

Our tokenizer is not suitable for tasks requiring precise semantic understanding, like VQA, because the MLC is trained in an unsupervised manner. Without additional constraints, such as contrastive learning between image features from Qwen-VLM and text embeddings in our codebook, aligning the image and text domains is challenging. A promising direction for future work is to enhance SweetTok into a semantically strong tokenizer by contrastive learning.

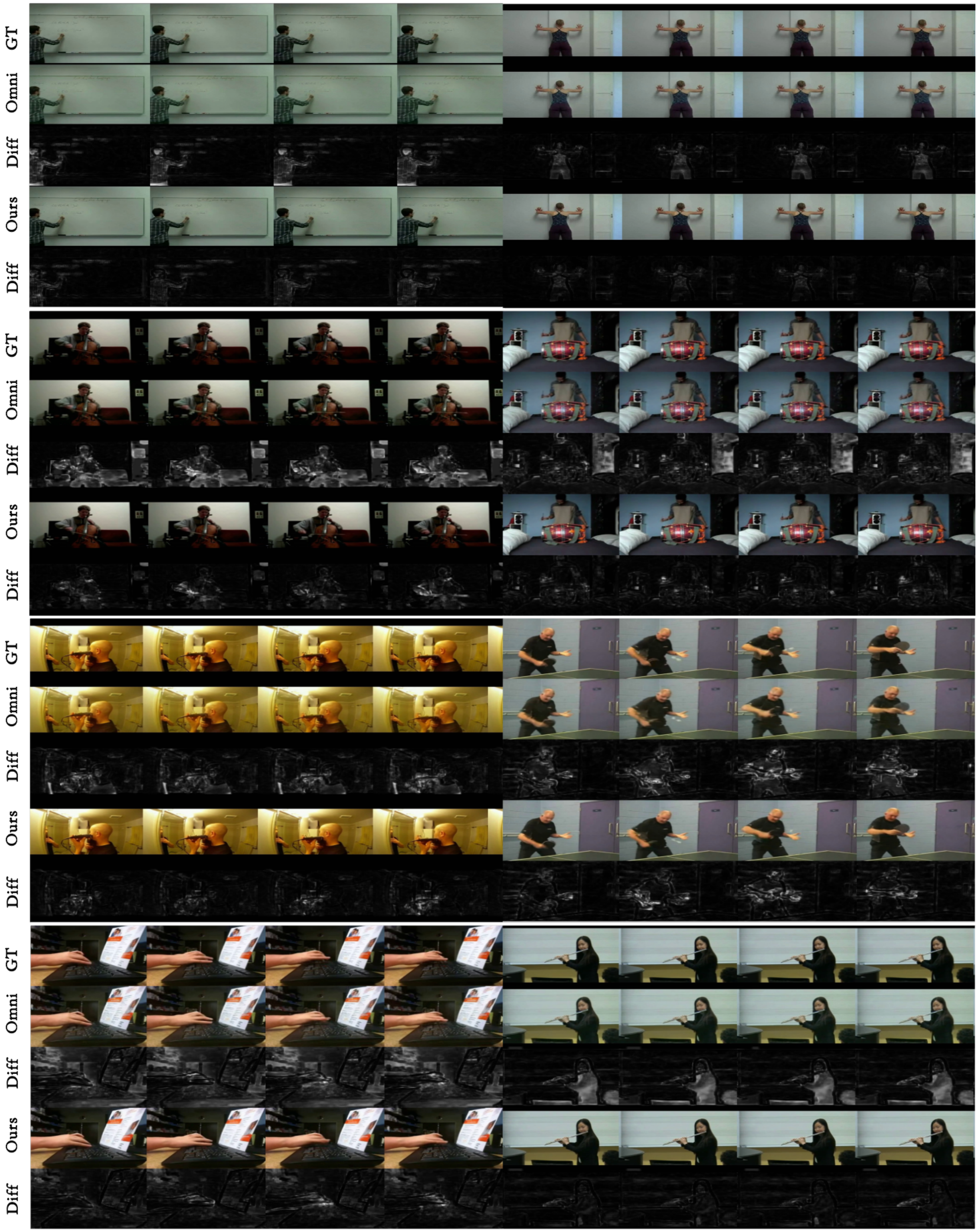

Figure 1. Comparison of the reconstruction results of OmniTokenizer and SweetTok on UCF-101 dataset, where ”Diff” represents the pixel difference between the ground truth and the models.

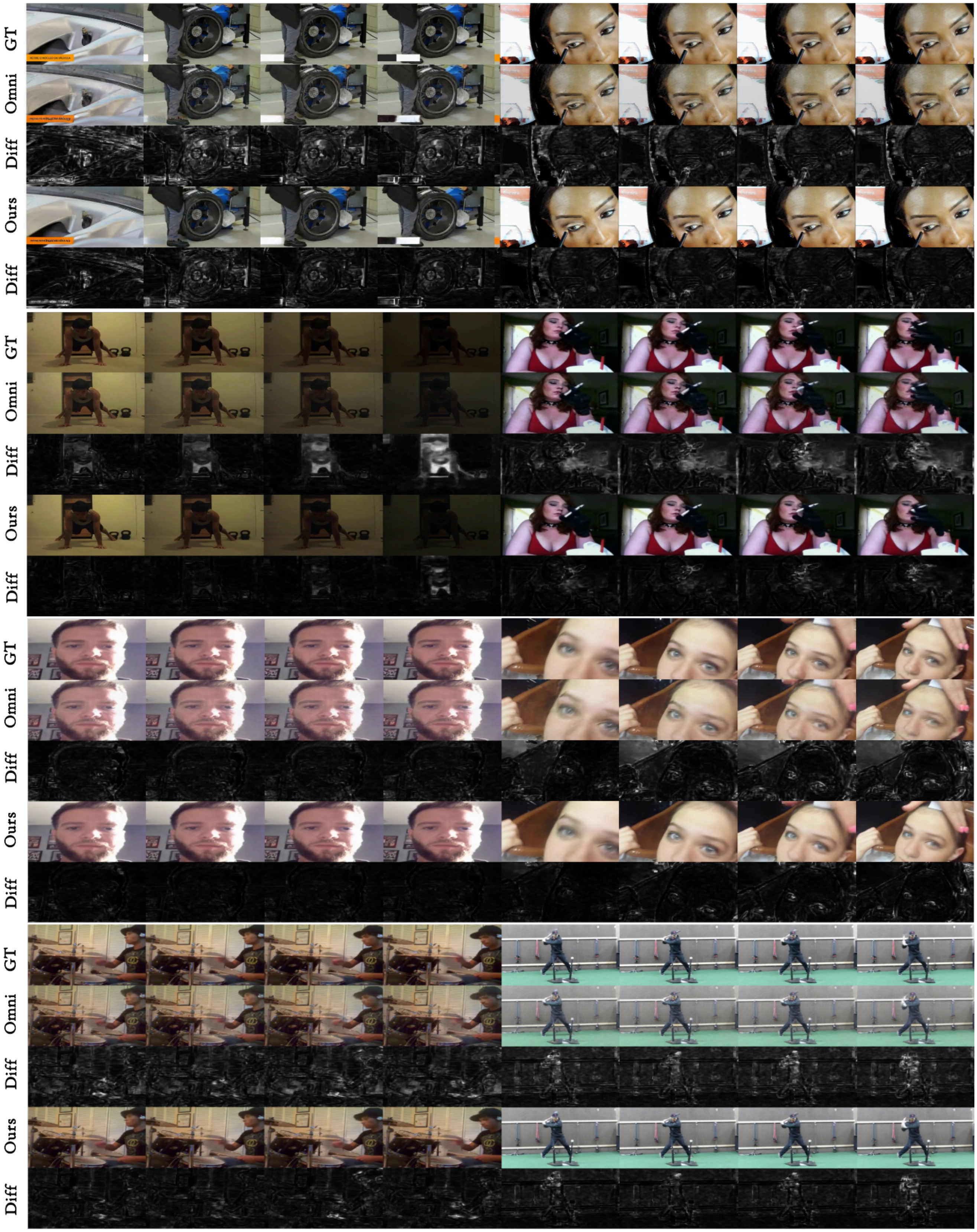

Figure 2. Comparison of the reconstruction results of OmniTokenizer and SweetTok on K-600 dataset, where ”Diff” represents the pixel difference between the ground truth and the models.

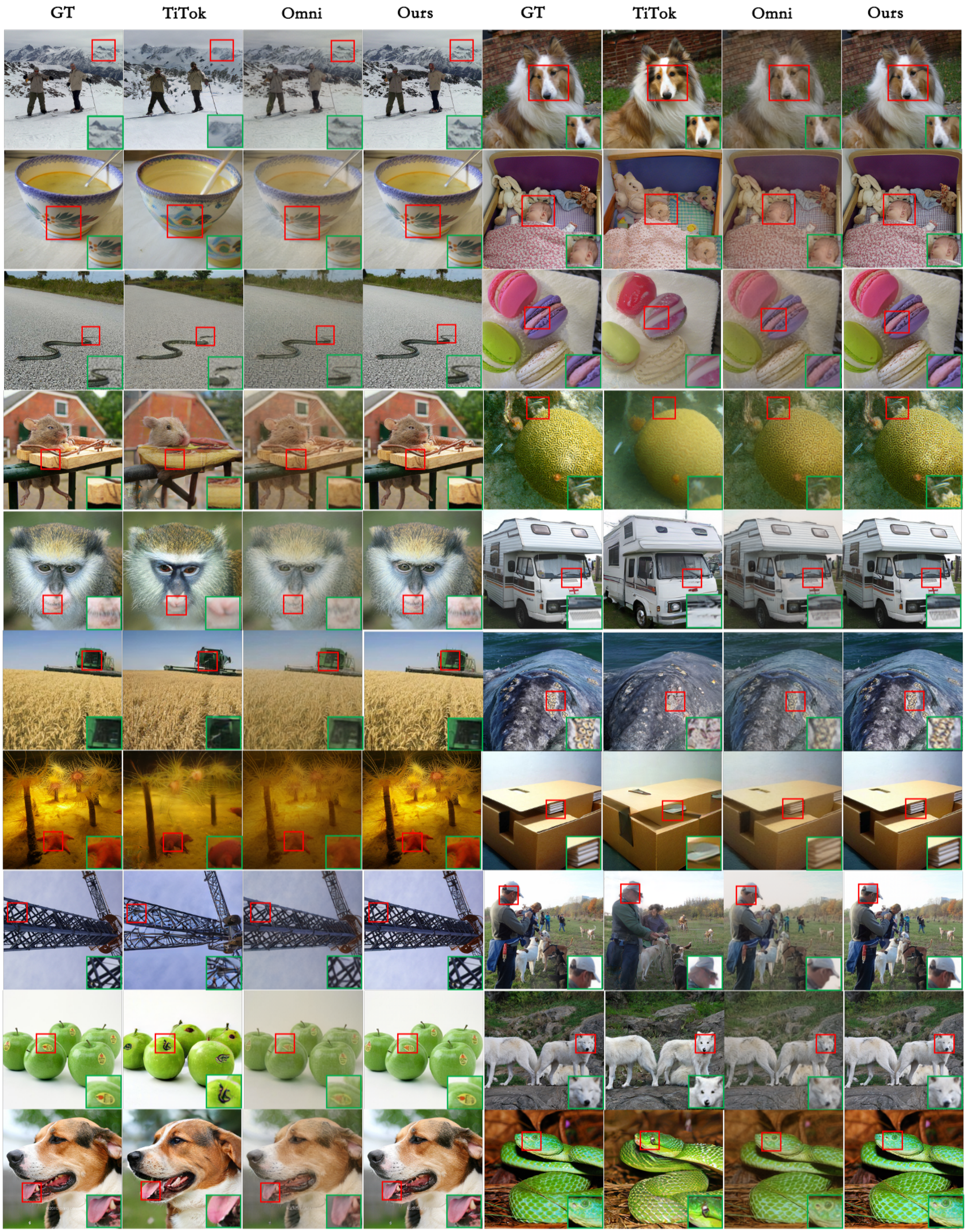

Figure 3. Comparison of the reconstruction results of TiTok, OmniTokenizer, and SweetTok on ImageNet-1K dataset. Differences are selected by the red blocks and highlighted in the grean blocks.

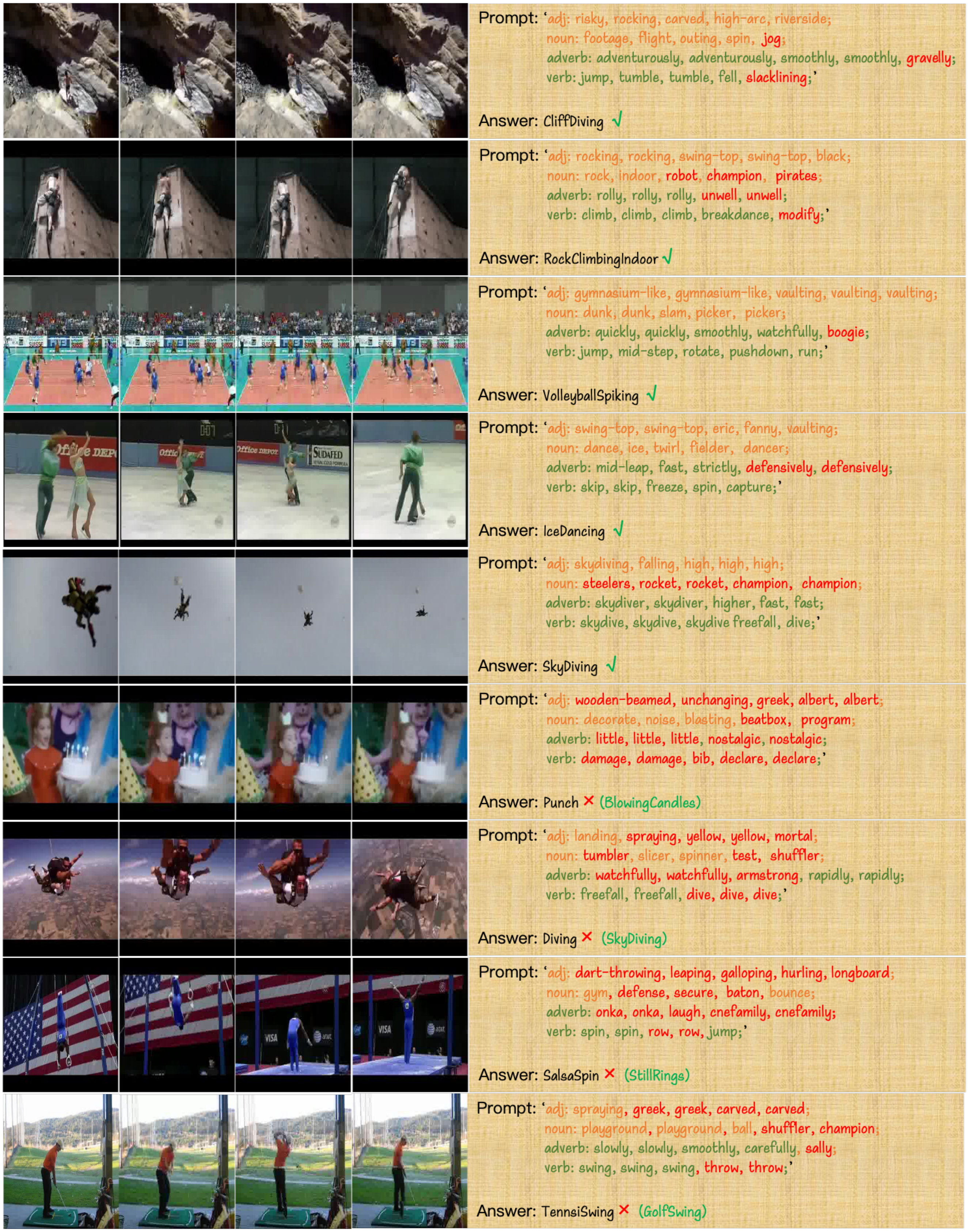
Prompt: 'adj: risky, rocking, carved, high-arc, riverside;
noun: footage, flight, outing, spin, jog;
adverb: adventurously, adventurously, smoothly, smoothly, gravelly;
verb: jump, tumble, tumble, fell, slacklining;'
Answer: CliffDiving √
Prompt: 'adj: rocking, rocking, swing-top, swing-top, black;
noun: rock, indoor, robot, champion, pirates;
adverb: rolly, rolly, rolly, unwell, unwell;
verb: climb, climb, climb, breakdance, modify;'
Answer: RockClimbingIndoor √
Prompt: 'adj: gymnasium-like, gymnasium-like, vaulting, vaulting, vaulting;
noun: dunk, dunk, slam, picker, picker;
adverb: quickly, quickly, smoothly, watchfully, boogie;
verb: jump, mid-step, rotate, pushdown, run;'
Answer: VolleyballSpiking √
Prompt: 'adj: swing-top, swing-top, eric, fanny, vaulting;
noun: dance, ice, twirl, fielder, dancer;
adverb: mid-leap, fast, strictly, defensively, defensively;
verb: skip, skip, freeze, spin, capture;'
Answer: IceDancing √
Prompt: 'adj: skydiving, falling, high, high, high;
noun: steelers, rocket, rocket, champion, champion;
adverb: skydiver, skydiver, higher, fast, fast;
verb: skydive, skydive, skydive freefall, dive;'
Answer: SkyDiving √
Prompt: 'adj: wooden-beamed, unchanging, greek, albert, albert;
noun: decorate, noise, blasting, beatbox, program;
adverb: little, little, little, nostalgic, nostalgic;
verb: damage, damage, bib, declare, declare;'
Answer: Punch ✗ (BlowingCandles)
Prompt: 'adj: landing, spraying, yellow, yellow, mortal;
noun: tumbler, slicer, spinner, test, shuffler;
adverb: watchfully, watchfully, armstrong, rapidly, rapidly;
verb: freefall, freefall, dive, dive, dive;'
Answer: Diving ✗ (SkyDiving)
Prompt: 'adj: dart-throwing, leaping, galloping, hurling, longboard;
noun: gym, defense, secure, baton, bounce;
adverb: onka, onka, laugh, cnefamily, cnefamily;
verb: spin, spin, row, row, jump;'
Answer: SalsaSpin ✗ (StillRings)
Prompt: 'adj: spraying, greek, greek, carved, carved;
noun: playground, playground, ball, shuffler, champion;
adverb: slowly, slowly, smoothly, carefully, sally;
verb: swing, swing, swing, throw, throw;'
Answer: TennsiSwing ✗ (GolfSwing)

Figure 4. Semantic words visualization for UCF-101. The visualization is based on few shot video action recognition tasks.

References

[1] Omnitokenizer: A joint image-video tokenizer for visual generation. *NeurIPS*, 2024. 1

[2] Larp: Tokenizing videos with a learned autoregressive generative prior. *ICLR Oral*, 2025. 1

[3] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 1

[4] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1

[5] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *NeurIPS*, 2024. 1, 2