# Single-step Diffusion-based Video Coding with Semantic-Temporal Guidance

Naifu Xue[1], Zhaoyang jia[2], Jiahao Li[3], Bin Li[3], Zihan Zheng[2], Yuan Zhang[1], Yan Lu[3]

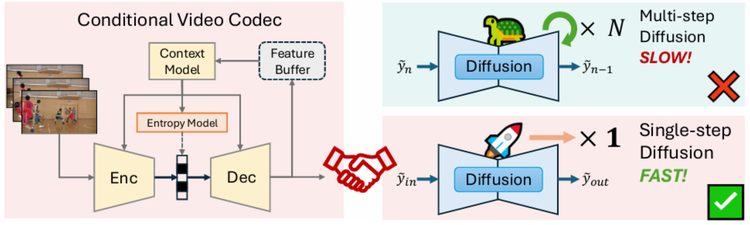
[1]Communication University of China, [2]University of Science and Technology of China, [3]Microsoft Research Asia

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Introduction

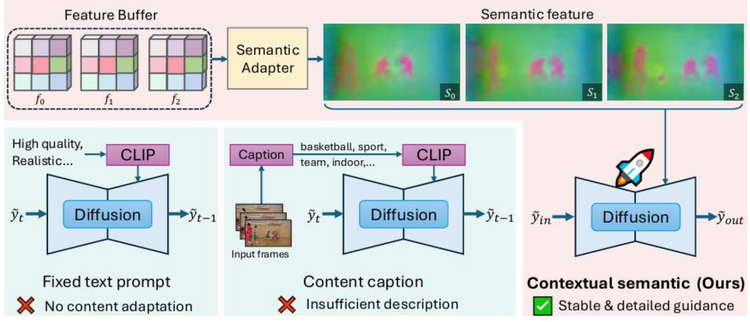
**Motivation:** Existing codecs fail to produce high perceptual-quality video reconstruction at **ultra-low bitrates.**

- Traditional codecs produce blocking, ringing, or blurry reconstructions.
- Previous generative codecs (e.g., GAN-based) produce noticeable artifacts.
- Diffusion-based codecs suffer from slow multi-step sampling.
- **Video requires both realism and temporal consistency.**



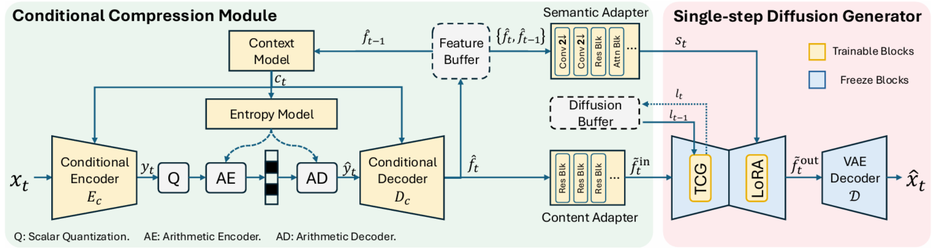

*Fast decoding + high-generation quality + temporal consistency?* **Our Solution: S²VC**

- How to improve the generation speed?
  - ➢ We integrate a distilled **single-step diffusion** into the conditional coding pipeline.
- How to enhance the generation quality?
  - ➢ We recognize the importance of semantic guidance for single-step diffusion and propose **Contextual Semantic Guidance (CSG)** to replace text-based conditioning.
- How to ensure the temporal consistency?
  - ➢ We propose **Temporal Consistency Guidance (TCG)**, adapting an image diffusion model into a temporally consistent video generator.



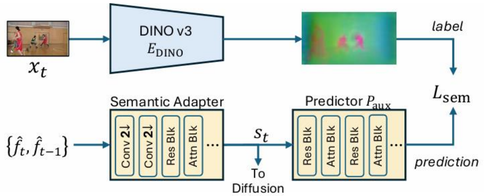

## Method

### Overall framework



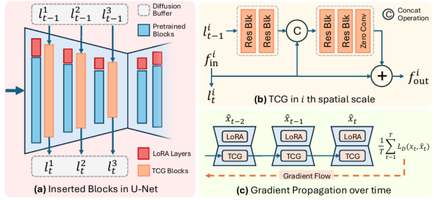

- **Conditional compression module** extracts compact context-aware representations for coding.
- **Single-step diffusion generator** restores realistic details with efficient one-step sampling.
- **Contextual Semantic Guidance (CSG)** provides temporally stable semantic conditioning from buffered features instead of text prompts.
- **Temporal Consistency Guidance (TCG)** propagates diffusion features across frames to ensure temporally coherent reconstruction.

### Contextual Semantic Guidance (CSG)




- Extracts frame-adaptive semantic guidance from buffered codec features.
- Replaces text prompts/captions with content-aware conditioning for single-step diffusion.
- Distills temporally stable semantics from DINOv3 via $L_{sem}$, with no extra inference cost.

### Temporal Consistency Guidance (TCG)




- Inserts plug-in TCG blocks into multiple U-Net scales to introduce temporal modeling.
- Reuses previous-frame diffusion features and fuses them with current-frame features.
- Cascade training propagates gradients across frames, improving temporal consistency.

### Training Objective

$$L = \frac{1}{T}\sum_{t=1}^{T}(\lambda R + L_{\mathrm{D}} + \alpha L_{\mathrm{sem}} + \beta L_{\mathrm{motion}})$$

Where:

$$L_{\mathrm{sem}} = ||E_{\mathrm{DINO}}(x_t) - P_{\mathrm{aux}}(s_t)||_1$$

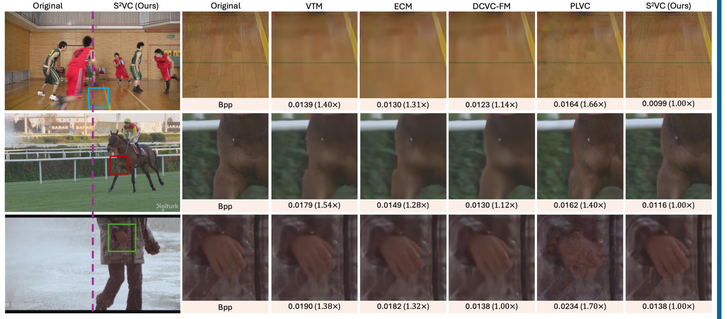
$$L_{\mathrm{D}} = ||x_t - \hat{x}_t||_1 + L_{\mathrm{LPIPS}}(x_t, \hat{x}_t)$$

$$L_{\mathrm{motion}} = ||O(x_{t-1}, x_t) - O(\hat{x}_{t-1}, \hat{x}_t)||_1$$

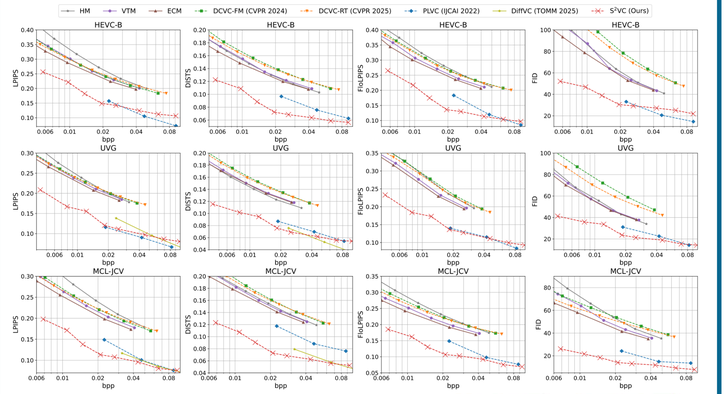
- $R$: bitrate estimated by entropy model
- $L_D$: distortion loss with L1 and LPIPS term.
- $L_{sem}$: semantic distillation loss
- $L_{motion}$: optical flow–based motion loss

## Experiment

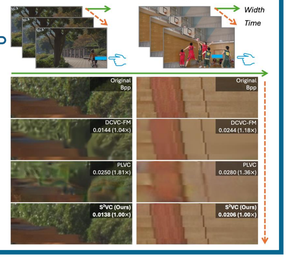
### Qualitative Evaluation: S²VC delivers best quality with lowest bitrate.




### Quantitative Evaluation: S²VC achieves SOTA rate-distortion result.




**Comparison of temporal consistency** by stacking the blue line across frames. The misaligned lines is caused by camera motion. Results closer to the original imply better temporal consistency and visual quality.




**Ablation studies** validate the effectiveness of the CSG and TCG.

| Dataset | HEVC-B | | | | UVG | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics for BD-Rate Calculation | LPIPS | DISTS | FloLPIPS | FID | LPIPS | DISTS | FloLPIPS | FID |
| *Contextual Semantic Guidance (CSG)* | | | | | | | | |
| w/o CSG | 27.46 | 22.08 | 25.65 | 28.05 | 43.75 | 32.16 | 39.32 | 18.36 |
| w/ CSG only | 13.30 | 14.06 | 14.10 | 7.28 | 34.00 | 29.71 | 28.38 | 13.97 |
| w/ CSG + distill → **Ours** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Temporal Consistency Guidance (TCG)* | | | | | | | | |
| w/o TCG | 20.41 | 23.67 | 28.13 | 9.54 | 38.75 | 26.70 | 46.88 | 4.87 |
| w/ TCG only | 11.74 | 11.25 | 14.66 | 6.32 | 33.92 | 21.92 | 28.46 | 1.96 |
| w/ TCG + cascade → **Ours** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |