

Free-VSC: Free Semantics from Visual Foundation Models for Unsupervised Video Semantic Compression

Yuan Tian^{1,2} , Guo Lu¹ , and Guangtao Zhai¹ 

¹ Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

² Shanghai Artificial Intelligence Laboratory

tianyuan168326@outlook.com, { luguo2014, zhaiguangtao}@sjtu.edu.cn


Abstract. Unsupervised video semantic compression (UVSC), i.e., compressing videos to better support various analysis tasks, has recently garnered attention. However, the semantic richness of previous methods remains limited, due to the single semantic learning objective, limited training data, etc. To address this, we propose to boost the UVSC task by absorbing the off-the-shelf rich semantics from VFMs. Specifically, we introduce a VFMs-shared semantic alignment layer, complemented by VFM-specific prompts, to flexibly align semantics between the compressed video and various VFMs. This allows different VFMs to collaboratively build a mutually-enhanced semantic space, guiding the learning of the compression model. Moreover, we introduce a dynamic trajectory-based inter-frame compression scheme, which first estimates the semantic trajectory based on the historical content, and then traverses along the trajectory to predict the future semantics as the coding context. This reduces the overall bitcost of the system, further improving the compression efficiency. Our approach outperforms previous coding methods on three mainstream tasks and six datasets.

Keywords: Video Compression · Video Analysis · Visual Foundation Models

1 Introduction

Video compression methods including traditional [7, 82] and learnable [48, 58, 59, 68] ones, have achieved remarkable advances in terms of visual quality metrics such as PSNR. However, directly deploying these methods to downstream analysis tasks usually achieves undesirable performance [113], due to not particularly preserving semantics during compression, which calls for efforts on video semantic compression.

The previous works can be roughly divided into supervised and unsupervised ones. Supervised methods, as shown in Figure 1 (b), are typically coupled with target tasks, optimizing the codec with task models [25, 51] together. The codec has to be adapted for the new task by a time-consuming supervised-training procedure, before being deployed. More seriously, this usually leads to the performance degradation on other tasks. The problems above limit their practicality. Although there are some image-oriented methods [17, 38] that enable a multiple tasks-shared encoder, their decoder part shall

 Corresponding authors

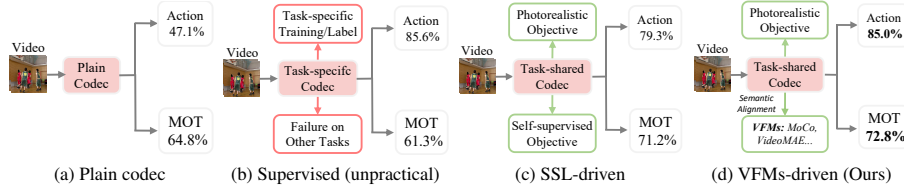


Fig. 1: Recent video semantic compression paradigms, where (a) deploying plain codecs like VVC directly to downstream tasks yields inferior results, (b) Adapting learnable codecs with task-specific supervisions, such as action recognition, requires pre-deployment training and data labels, yet performs poorly on other tasks like multiple object tracking (MOT), thus unpractical, (c) prior unsupervised methods driven by self-supervised learning (SSL) exhibit undesirable results due to the limited learned semantics, and (d) our approach, driven by pre-trained visual foundation models (VFMs), achieves strong results by absorbing their rich semantics. Evaluation is conducted on UCF101@0.02bpp and MOT17@0.01bpp, utilizing TSM [64] and ByteTrack [121], respectively. Network architectures of (b)/(c)/(d) are consistent for a fair comparison.

still be tuned towards the target task before deployment. The tuning procedure is non-trivial, considering that video annotations are laborious, and jointly fine-tuning codecs with video analysis models consumes heavy computational cost.

Therefore, recently, there emerges the unsupervised video semantic compression paradigm [89], as shown in Figure 1 (c), aiming to produce low-bitrate videos that can support various video analysis tasks. This paradigm optimizes a trade-off objective between the bitcost and the semantic distortion of the compressed video, along with a photorealistic term for regularizing the video visual quality. The primary challenge lies in representing the semantic content of videos, so that a proxy function of semantic distortion can be quantitatively calculated. Some early methods representing video semantics through hand-crafted representations, such as region-of-interest (ROI) [8, 39] or segmentation maps [3, 33], unable to enjoy the benefits of big data. Recently, a few approaches [36, 89] have integrated self-supervised learning objectives into compression. Although data-driven, these methods still face limitations in learning rich semantics, due to the adoption of simple and single learning objectives, limited training data, etc.

On the other hand, there has been a rapid advancement in the community of open-sourced Visual Foundation Models (VFMs) [43, 44, 94, 98]. These substantial-parameter models are trained with diverse learning objectives and large-scale datasets, probably emerging strong and complementary semantic representation capabilities [11]. This prompts a question: *can we harness the rich semantic representation within VFMs for the unsupervised video semantic compression problem?*

Our answer is YES. We propose a new semantic coding framework termed Free-VSC to absorb free and rich semantics from multiple VFMs. However, it is non-trivial to perform learning from several large-scale foundation models, considering the heterogeneity among their learned semantics. To cope with this, we introduce a prompt-based semantic alignment layer (Prom-SAL) to align the compressed video semantics to a shared representation space of various foundation models, as shown in Figure 1 (d). The parameters of Prom-SAL are shared for all VFMs, implicitly building a mutually-enhanced and harmonious semantic space, while the input prompts to Prom-SAL are VFM-specific and modulate the alignment procedure to adapt to different VFMs.

As for the compression of inter-frame semantic redundancies, most previous compression methods [8, 51, 89] adopt the motion-based schemes. However, unlike low-level pixels/features that show obvious motions (spatial displacements), the temporal patterns of high-level semantic features are implicit and highly-related to the video content. Thus, we propose to predict the next frame by first predicting the video-adaptive semantic trajectory, instead of relying on explicit motion. The trajectories are modeled as distributions of non-linear transformations, which are dynamically predicted based on the historical information. Then, we transform the previous features along the trajectory to obtain a preliminary future prediction, substantially reducing the uncertainty of the future frame, as well as its bitcost [78]. Compared to previous motion or static transformation-based ones, the proposed scheme is more effective for semantic compression, due to its content-adaptiveness and flexible handling of high-level features. Our contributions are:

1. We propose a new video semantic compression framework by harnessing powerful Visual Foundation Models (VFMs), reusing their rich and free semantic representations. This is the first work employing VFMs for semantic compression.
2. A prompt-based semantic alignment layer is devised to facilitate the framework learning mutually-enhanced semantics from multiple VFMs, effectively guiding the compression model to particularly preserve the semantics within videos.
3. A dynamic trajectory-based entropy model is introduced to remove inter-frame semantic redundancy, by first predicting video-adaptive trajectories within semantic space, providing better semantic compression efficiency than other widely-used schemes.
4. Our approach demonstrates substantial improvements over previous methods on three tasks and six datasets.

2 Related Works

Image/Video Semantic Compression. Traditional [7, 82, 102, 115], learnable codecs [5, 48, 58, 59, 68, 69, 118, 120], and some tradition-neural mixed codecs [86, 91] are designed for achieving good visual quality, instead of preserving video semantics. To address this, early video coding for machine (VCM) standards such as CDVA [30] and CDVS [29, 31] transmit image keypoints, serving image indexing and retrieval tasks. Later, several methods [19, 20, 24, 38, 63, 80, 119] propose to compress intermediate feature maps instead of images, where the downstream models shall to be trained for adapting the compressed features. Additionally, various researches [8, 23, 26, 50, 116] have elevated traditional codecs by introducing an independent task-specific feature encoding stream [13, 96], or utilizing hand-crafted structure maps [3, 33, 46]. Furthermore, some works [4, 25, 32, 40, 112] directly incorporate the downstream task loss. Recently, several works [34, 38] have adopted self-supervised loss as the data-driven semantic source. But, they still require task-specific training, limiting their practicality.

Unsupervised Video Semantic Compression (UVSC). Recently, the concept of UVSC has emerged for catering to real-world scenarios, where compressed videos from a unified codec are directly used for various downstream tasks [6, 9, 14, 15, 21, 22, 28, 60, 61, 64, 84, 85, 90, 92, 97, 99, 100, 106, 106, 107, 114, 121] without any task-specific training. Recent works [87–89] propose to optimize a semantics-preserving objective such as contrastive learning [44] or MAE [43] [94] loss, as well as another photo-realism regularization item such as GAN [41], to compress the video into a semantic-

full yet high-quality video at very low bitrate. Nonetheless, the semantics obtained by a single self-supervised learning loss is still not sufficiently rich. In this work, we leverage the pre-trained VFMs to absorb much richer semantics.

Visual Foundation Models. Recently, numerous works have combined the powerful modeling capabilities of Transformers [27, 67, 95, 122] with self-supervised learning methods, resulting in a series of powerful VFMs for both images [43, 73, 104] and videos [94, 98, 100]. For the first time, we explore how to harness current VFMs for the unsupervised video semantic compression problem.

Inter-Frame Entropy Models. Traditional [7, 82] and state-of-the-art learnable codecs [47, 59] typically use motion to model the inter-frame redundancy. However, explicit pixel (or region) movement-based motions are often suitable for modeling the temporal dynamics of low-level features, instead of more abstract high-level semantic features. While a few methods [66, 70, 110] directly predict entropy model parameters or features for current frame, their models are static. In contrast, our trajectory-based entropy model is content-adaptive and dynamic.

3 Approach

In this paper, we propose a new unsupervised video semantic coding framework termed Free-VSC (Figure 2), which absorbs rich semantic knowledge from multiple VFMs. The main learning objective is to minimize the discrepancy between the compressed and original videos in a VFMs-aligned semantic space. Besides, a semantic trajectory-based entropy model is proposed for more efficient inter-frame semantic compression.

Encoding. Let an input video $X = \{x_1, x_2, \dots, x_t, \dots, x_T\}$, where T denotes the video temporal length, each frame x_t is first encoded and quantized as feature \hat{f}_t by a frame encoder on the sender side. Then, based on the historical feature information $\{\hat{f}_1, \dots, \hat{f}_{t-1}\}$, the possible semantic trajectory of each region is predicted as a distribution of non-linear transformations. The past frame feature \hat{f}_{t-1} is transformed along the trajectories, serving as the context feature for calculating the entropy model $p(\hat{f}_t)$. Finally, an arithmetic encoder (AE) uses $p(\hat{f}_t)$ to encode \hat{f}_t as bitcodes.

Decoding. On the recipient side, all past quantized features $\{\hat{f}_1, \dots, \hat{f}_{t-1}\}$ have been received and are accessible, so that the entropy model $p(\hat{f}_t)$ can be precisely reconstructed without transporting any side information. The bitcodes can be decoded back into \hat{f}_t with $p(\hat{f}_t)$ by an arithmetic decoder (AD), which are further restored as a video frame \hat{x}_t by a frame decoder network. The decoded frames are fed into various video analysis modules, such as video action recognition (ActR), multiple object tracking (MOT) and video object segmentation (VOS) models.

Learning. During the optimization of the compression system, we randomly select one VFM V_n from the candidate VFM set as the semantic guidance in each optimization step. Compared to simultaneously training with multiple VFMs, randomly activating a single VFM enables scaling VFM number more easily, due to the substantially saved GPU memory. Upon receiving the compressed video semantic feature $\hat{f} = \{\hat{f}_1, \dots, \hat{f}_T\}$, we employ a prompt-based semantic alignment layer (Prom-SAL) and the VFM V_n -specific prompt P_n to align it within the semantic space of V_n , thereby generating the VFM-aligned compressed semantic feature g_n . Then, a simple ℓ_2 loss is utilized to minimize the discrepancy between g_n and the original video feature extracted by V_n .

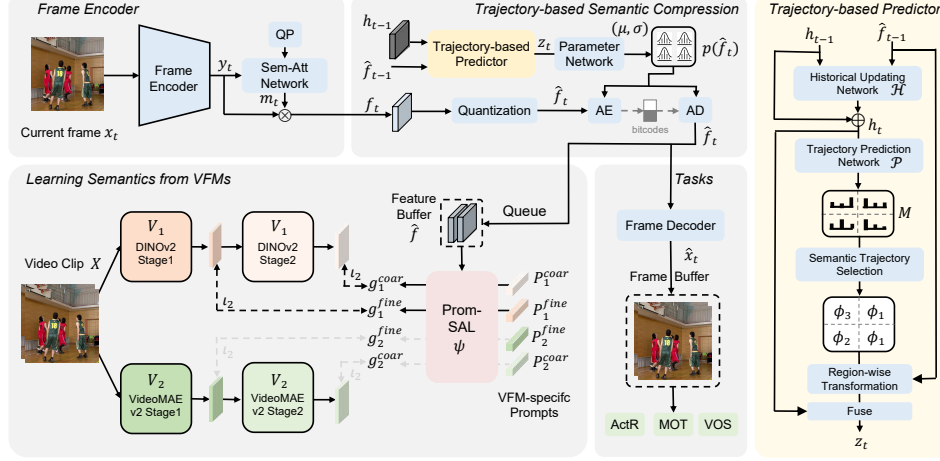


Fig. 2: Overview of Free-VSC framework, which learns to absorb rich semantics from multiple VFMs into the compression procedure. A prompt-based semantic alignment layer (Prom-SAL) is introduced to flexibly align the compressed video feature \hat{f} to the semantic space of VFMs. We also propose a trajectory-based entropy model for efficiently compressing the inter-frame semantic redundancy. We illustrate two VFMs (V_1 and V_2) and three semantic trajectories for simplicity, although more VFMs and trajectories can be applied in our approach.

The learning procedure is conducted in a single stage, *i.e.*, both the encoder, decoder and Prom-SAL are simultaneously optimized in an end-to-end fashion.

The details of each component are elaborated as follows.

3.1 Frame Encoder and Decoder Networks

We employ a frame encoder network to transform the frame $x_t \in \mathbb{R}^{3 \cdot H \cdot W}$ into deep features $y_t \in \mathbb{R}^{256 \cdot \frac{H}{32} \cdot \frac{W}{32}}$. To adapt to various compression quality parameters (QPs), we further append a QP-conditioned semantic attention network (Sem-Att Net). This network generates a spatial-channel attention mask m_t for masking the information according to different QP values. The final semantic feature is obtained by multiplying the original feature with the attention mask, denoted as $f_t = y_t \odot m_t \in \mathbb{R}^{256 \cdot \frac{H}{32} \cdot \frac{W}{32}}$, where \odot represents the element-wise multiplication operation. The resulting feature f_t is quantized as \hat{f}_t for transmission. We mention that using QP as the conditional variable to adjust feature bitrate is a common practice in variable-rate compression [108].

The encoder frame network consists of five stages. The initial stage includes a convolution layer with stride size two, followed by a residual block. Each of the subsequent four stages comprises a convolution layer with stride size two, succeeded by two inverted residual blocks [77] and two residual blocks. The Sem-Att Net is simply composed of three convolution layers with kernel size three, followed by a sigmoid function.

The frame decoder comprises five stages, progressively upsampling low-resolution features \hat{f}_t to full-resolution frames. The first stage includes a single upsampling layer implemented with transpose convolution. The second/third stages include eight/four residual blocks followed by an upsampling layer, while the last two stages include only one residual block for efficient computation.

3.2 Trajectory-based Semantic Compression

As shown in the right side of Figure 2, we first recurrently aggregate the past information as the hidden state h_t , which is of the same size as the frame feature f_t . Then, we predict the context feature z_t of the current frame feature \hat{f}_t via dynamically predicting the region-wise semantic trajectories. Finally, conditioned on z_t , a contextual entropy model is employed to compress \hat{f}_t . The detailed procedure is expanded as follows.

Historical State Aggregation. Formally, given the previously aggregated state h_{t-1} and the previous frame feature \hat{f}_{t-1} , we employ a historical updating network \mathcal{H} to combine them, resulting in a new state $h_t = h_{t-1} + \mathcal{H}(h_{t-1}, \hat{f}_{t-1})$. Subsequently, h_t is recurrently used as the historical information in the coding loop of the next frame. The network \mathcal{H} comprises three convolutions with kernel size three and LeakyReLUs [105].

Trajectory-based Context Prediction. Rather than directly predicting context from h_t using a static transformation, which can struggle with diverse contents across different frames or videos, we propose to (1) first dynamically predict the possible semantic trajectory of each region, and (2) then synthesize the context by transforming the previous feature with trajectories in a region-wise manner. We select the dynamic trajectories from a few possible candidates instead of directly predicting their parameters for easier optimization, following previous work [18]. The possible trajectories are defined as a pool of non-linear transformations $\{\phi_1, \phi_2, \dots, \phi_S\}$, where S denotes the total number.

We employ a trajectory prediction network to predict the region-wise trajectory probability map M_t . For each spatial position $o = \{o_x, o_y\}$, the transformation with the largest probability is selected as the semantic trajectory ϕ_s . ϕ_s is used to transform the previous feature $\hat{f}_{t-1}[o]$ to predict the future feature $f'_t[o] = \phi_s(\hat{f}_{t-1}[o])$. Finally, we utilize a Fuse operation to enhance the predicted future feature f'_t with the long-term information, producing the current frame coding context, *i.e.*, $z_t = \text{Fuse}(f'_t \oplus h_t)$, where \oplus denotes the channel-wise concatenation operation.

As for network details, the trajectory prediction network consists of three convolution layers with kernel size three, followed by a *softmax* operation. Fuse comprises three residual blocks, each of which includes one convolution with kernel size three.

Context-based Compression. Given the context feature z_t , we employ three stacked convolutions with kernel size three, followed by a three-layer multiple-layer-perceptron (MLP), to predict two sets of parameters, namely mean μ and variance σ , which are used to construct the Gaussian distribution model $p(\hat{f}_t) \sim \mathcal{N}(\mu, \sigma)$. \hat{f}_t is compressed by an arithmetic encoder with $p(\hat{f}_t)$ [5]. The sizes of μ and σ are the same as that of \hat{f}_t .

3.3 Learning Rich Semantics from VFMs

Given the compressed video semantic feature $\hat{f} = \{\hat{f}_1, \dots, \hat{f}_T\}$, we propose to align it with pre-trained VFMs, within the shared semantic space of VFMs. We compute the feature distance in this aligned space as the semantic distortion term. In each iteration, a VFM is randomly selected as the semantic source, avoiding balancing different VFMs. Additionally, this approach involves only one VFM in each forward-backward procedure, ensuring memory efficiency. In the following, we provide a detailed description of the semantic learning procedure.

Given the selected VFM V_n , we input the original video X into it, resulting in two groups of feature maps representing fine-level semantics (in shallow layer) and coarse-level semantics (in deep layer), denoted as, $V_n(X)^{fine}$ and $V_n(X)^{coar}$. The adoption

of multi-level features is for catering to the downstream tasks of various granularities, *i.e.*, global tasks such as recognition rely mostly on coarse-level features, while local tasks such as tracking/segmentation require both coarse- and fine-level features. These two-level features are separately normalized by dividing the moving average of their ℓ_2 norm. This normalization step eliminates the magnitude differences among features obtained from different VFMs. For simplifying the notations, we denote the normalized features also as $V_n(X)^{fine}$ and $V_n(X)^{coar}$.

Then, we introduce the prompt-based semantic alignment layer (Prom-SAL) to align our compressed semantics to the semantic features from different VFMs. Prom-SAL mainly includes a spatial-temporal Transformer ψ . Motivated by the conclusion in VPT [54], *i.e.*, input prompts can effectively tune Transformer’s behavior towards a specific learning target, we also equip ψ with a set of the VFM-specific prompt tokens. We denote the prompts for V_n as $\{P_n^{fine} \in \mathbb{R}^{1024}, P_n^{coar} \in \mathbb{R}^{1024}\}$, which are learnable and initialized with values drawn from a Gaussian distribution. Given the inputs above, Prom-SAL produces the VFM V_n -aligned features $g_n^{fine} = \psi(P_n^{fine}, \hat{f})$ and $g_n^{coar} = \psi(P_n^{coar}, \hat{f})$. Then, the semantic distortion can be formulated as:

$$\mathcal{D}_{sem} = \ell_2(g_n^{fine}, V_n(X)^{fine}) + \ell_2(g_n^{coar}, V_n(X)^{coar}), \quad (1)$$

where ℓ_2 denotes the mean square error (MSE) loss.

Finally, the trade-off objective between the bitrate and the distortion of the semantic feature can be formulated as,

$$\mathcal{L}_{RDsem} = QP \cdot \frac{1}{T} \sum_{t=1}^T R(\hat{f}_t) + \mathcal{D}_{sem}, \quad (2)$$

where the compression quality parameter QP is randomly sampled from $[0.5, 4]$, aiming to train a unified compression model for variable bitrates. T denotes the length of the training clip. $R(\hat{f}_t)$ is the entropy [5] of \hat{f}_t with the distribution $p(\hat{f}_t)$. Following previous works [89] [35], we also enforce the image perceptual loss [117] \mathcal{L}_{percep} and Generative Adversarial Network (GAN) [41] loss \mathcal{L}_{GAN} on the decoded frame \hat{x}_t , to ensure the photo-realism of the decoded frames. Finally, the whole loss function of the system is given by $\mathcal{L} = \mathcal{L}_{RDsem} + \mathcal{L}_{percep} + \mathcal{L}_{GAN}$. The architecture and loss function of GAN follow that of PatchGAN [53].

4 Experiments

4.1 Experimental Setup

Training Dataset. Following the previous works [89, 103], we use 60K videos from the Kinetics400 dataset [12] during the training stage. For data augmentation, the video sequences are randomly flipped and cropped into 256×256 patches.

Evaluation Datasets. Our evaluation protocol rigorously follows [89]. For action recognition task, we evaluate it on UCF101 [81], HMDB51 [57], Kinetics [12], and Diving48 [62]. For multiple object tracking (MOT) task, we evaluate it on MOT17 [71]. For video object segmentation (VOS) task, we evaluate it on DAVIS2017 [76].

VFM Details. We have adopted three representative self-supervised VFMs: -VideoMAEv2 [98], utilizing ViT-B network [27] and trained with MAE loss [43]. -DINOv2 [73], based on ViT-B network and trained using a combined objective, incorporating DINO [11], iBOT [123], and SwAV [10]. -SSL-Swin [104], leveraging the Swin-B network [67]

and trained with a combined objective of MoCo v2 [16] and BYOL [42]. These models cover mainstream self-supervised learning paradigms, such as masked image/video modeling [43] and contrastive learning [44]. Additionally, they incorporate two popular backbone architectures: non-hierarchical (ViT) and hierarchical (Swin). The ‘fine’ and ‘coarse’ layers in all VFMs correspond to the ‘4th’ and ‘8th’ attention blocks. We avoid using very deep layers, as they are biased towards the used self-supervised learning objectives, potentially limiting generalizability.

Implementation Detail. The I frame is encoded with the channel-wise auto-regressive entropy model [72], where the channel group number is set to eight. For the first P frame, the historical state h_{t-1} of the entropy model is initialized as the I frame feature. The Transformer in Prom-SAL is instantiated as six attention blocks of ViT, where input channel number, MLP hidden channel number and attention header number are set to 1024, 2048 and 32, respectively. \sin/\cos positional embedding is adopted. The number of possible trajectories N is empirically set to 5. The quality parameter (QP) set is $\{0.5, 2, 3, 4\}$. The video clip length T , is fixed at eight during training. The initial learning rate is set to $1e-4$ and is reduced by a factor of ten at 500,000 steps. The total number of training steps is 600,000. The mini-batch size is 16. We utilize the Adam optimizer [56] implemented in PyTorch [75] with CUDA support. The values of β_1 and β_2 are set to 0.9 and 0.999, respectively. The entire training process takes approximately five days on a machine equipped with four NVIDIA GeForce RTX 4090 GPUs.

Baseline Methods. We compare our method with traditional codecs, namely HEVC [93] and VVC [7,101], learnable codecs, namely FVC [48], PLVC [111] and DCVC-DC [59], as well as recent semantics-oriented coding methods, namely ROI [8], JPD-SE [33] and SMC [89]. All approaches are evaluated on low delay P mode (LDP) with group-of-picture(GOP) size 10 for a fair comparison.

Task Models. For the action recognition task, we evaluate on TSM [64], SlowFast [37], and TimeSformer [6] networks. The model weights are provided by MMAction2 [2]. For the MOT task, we adopt ByteTrack [121]. The model weight is provided by MMTTracking [1]. For the VOS task, we adopt XMem [22] and the official model.

Evaluation Metrics. We use bpp (bit per pixel) to measure the average number of bits used for one pixel in each frame. For the action recognition task, we adopt the Top1 accuracy metric. For the MOT task, we adopt MOTA (multiple object tracking accuracy) [55], MOTP (multiple object tracking precision), FN (false negative detection number) and IDF1 metrics. For the VOS task, Jaccard index \mathcal{J} , contour accuracy \mathcal{F} , the average of \mathcal{J} and \mathcal{F} ($\mathcal{J}\&\mathcal{F}$), and contour recall $\mathcal{F}\text{-Recal}$ are adopted.

4.2 Experimental Results

Action Recognition. As shown in Table 1, when using TSM as the action recognition network, our method demonstrates superior performance to all other methods, *i.e.*, the recent learnable method DCVC-DC [83] and semantic codec SMC [89]. On lower bitrate settings, such as HMDB51@0.02bpp and Diving48@0.05bpp, we observe a 27.02% and 19.30% performance gain over DCVC-DC, along with 10.12% and 12.01% accuracy enhancements over SMC. In higher bitrate settings, our method also maintains superiority, *e.g.*, outperforming SMC by 2.70%, 2.57%, and 7.67% on UCF101@0.04bpp, Kinetics400@0.06bpp, and Diving48@0.07bpp settings, respectively. Moreover, when compared to other hand-crafted semantic representation-based

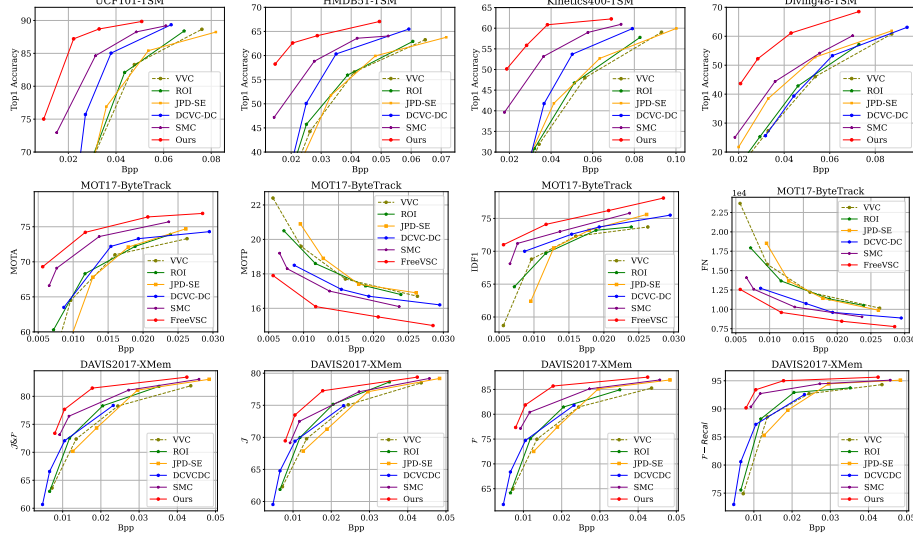


Fig. 3: Semantic compression performance on Action Recognition, MOT and VOS tasks. The plot titles are in {Dataset}-{Model} format.

Table 1: Results on action recognition in terms of Top1 accuracy (%). ‘Upper-bound’ is obtained by evaluating action recognition models with the original dataset. ‘K400’ denotes Kinetics400.

	TSM				Slowfast			TimeSformer	
	HMDB51	Diving48	UCF101	K400	Diving48	UCF101	HMDB51	UCF101	HMDB51
<i>Bpp</i>	@ 0.02	@ 0.05	@ 0.04	@ 0.06	@ 0.07	@ 0.03	@ 0.02	@ 0.04	@ 0.04
HEVC	33.66	22.48	52.70	35.28	37.24	85.13	47.78	69.92	37.45
FVC [48]	32.48	22.94	64.61	37.23	46.29	79.43	52.15	72.47	45.11
PLVC [111]	43.92	30.13	71.40	48.65	42.63	87.38	56.33	70.06	44.59
VVC	34.35	42.75	76.97	49.11	53.23	86.93	59.95	85.10	57.83
ROI [8]	35.81	44.94	78.63	49.82	55.64	87.25	60.44	85.91	58.21
JPD-SE [33]	38.76	50.27	78.85	51.02	57.22	84.86	60.22	84.03	57.79
DCVC-DC [59]	35.26	44.07	85.39	55.81	55.93	86.27	60.83	87.66	55.34
SMC [89]	52.12	51.36	86.46	59.26	60.16	88.89	63.48	89.60	60.74
Ours	62.24	63.37	89.16	61.83	67.83	90.53	65.42	92.16	64.87
Upper-bound	72.81	75.99	93.97	70.73	75.99	94.92	72.03	95.43	71.44

methods ROI [8] and JPD-SE [33], our approach achieves more than 10% performance gain on the large-scale Kinetics400@0.06bpp setting.

From the above comparison, we observe that an advanced video semantic compression system requires the particular preservation of semantics, which is lacking in plain codecs like VVC and DCVC-DC. Furthermore, semantics derived from VFMs are more effective than SSL-learned semantics in SMC and hand-crafted methods like ROI. Additionally, our approach shows remarkable improvements on the challenging Diving48 dataset, which requires reasoning about fine-grained temporal relationships. The significant progress can be attributed to the proposed inter-frame entropy model, which compresses video semantics by predicting temporal dynamics. Our approach is also effective

Table 2: MOT and VOS performances of different coding methods. \downarrow denotes the lower is better. ‘Upper-bound’ is obtained by evaluating the task models with the original datasets.

	MOT: ByteTrack on MOT17@0.01bpp				VOS: XMEM on DAVIS2017@0.01bpp			
	MOTA (%) \uparrow	MOTP (%) \downarrow	IDF1 (%) \uparrow	FN \downarrow	$\mathcal{J}\&\mathcal{F}$ (%)	\mathcal{J} (%)	\mathcal{F} (%)	\mathcal{F} -Recal (%)
HEVC	61.30	19.88	64.32	17377	57.68	56.84	58.51	67.44
FVC [48]	44.24	21.97	52.53	27508	62.39	61.22	63.55	75.67
PLVC [111]	67.87	18.44	68.95	13299	61.45	60.02	62.87	74.07
VVC	64.86	19.49	68.99	15621	67.47	65.59	69.36	80.92
ROI [8]	65.29	19.31	67.78	15270	69.22	67.16	71.28	83.80
JPD-SE [33]	60.05	20.62	63.52	17847	61.48	59.86	63.09	73.31
DCVC-DC [59]	65.22	18.22	70.51	12328	72.86	69.26	76.46	88.17
SMC [89]	70.84	17.79	71.89	11710	74.20	70.21	78.19	91.10
Ours	72.80	16.61	73.21	10445	76.85	72.70	80.99	92.77
Upper-bound	78.60	15.80	79.00	7000	87.70	84.06	91.33	97.02



Fig. 4: Qualitative comparison between the compressed frame by different methods. The frame is from HEVC Class C dataset. The numbers in parentheses indicates the compression ratio.

tive on more advanced action networks like Slowfast (+1.94% on HMDB51@0.02bpp) and TimeSformer (+2.56% on UCF101@0.04bpp).

Finally, we also compare different coding approaches on the fine-grained action recognition dataset FineGym99 [79]. Our approach (71.51% Top1 accuracy) substantially outperforms VVC (62.43%) and SMC (42.84%) by a larger margin at 0.02bpp bitrate level, when using TSM as the recognition network.

Multiple Object Tracking (MOT). In addition to the basic action recognition task, we also compare the coding methods on a much more challenging MOT task. As shown in the left part of Table 2, our method achieves the state-of-the-art performance in terms of all metrics. For example, our approach outperforms DCVC-DC and SMC by 7.58% and 1.96% on the MOT17@0.01bpp setting, respectively, in terms of MOTA. Finally, we provide the RP curves on MOT task in the second row of Figure 3.

Video Object Segmentation (VOS). We further compare our method with previous ones on VOS task, which requires more fine-grained semantic features than the above two tasks. As shown in the right part of Table 2, our method achieves the best performance in terms of all metrics. For example, our approach outperforms DCVC-DC and SMC by 3.99% and 2.65% on the DAVIS2017@0.01bpp setting, respectively, in terms of $\mathcal{J}\&\mathcal{F}$. We also illustrate the RP curves on VOS task in the third row of Figure 3.

Visual Quality. As shown in Figure 4, when compressing the videos by over 1300 times, our method shows better visual quality than DCVC-DC and SMC, *e.g.*, sharper edges or less checkerboard artifacts, while with larger compression ratio (1519x). We also recruit four volunteers to score semantic completeness of the compressed frames from HEVC Class C dataset [82] by different coding methods. The score range is from one to five, with higher scores indicating that objects and events within videos are more easily identifiable. As shown in Figure 4 (e), our method achieves remarkably higher

mean opinion score (MOS) [52]. Furthermore, our method achieves lower LPIPS [117] than SMC by about 0.05 on average across all bitrate levels.

4.3 Ablation Study and Analysis

To validate each design of the proposed framework, we perform thorough ablation studies. For simplicity, numerical results in tables and text are reported based on the UCF101@0.02bpp setting using the TSM action network, unless stated otherwise.

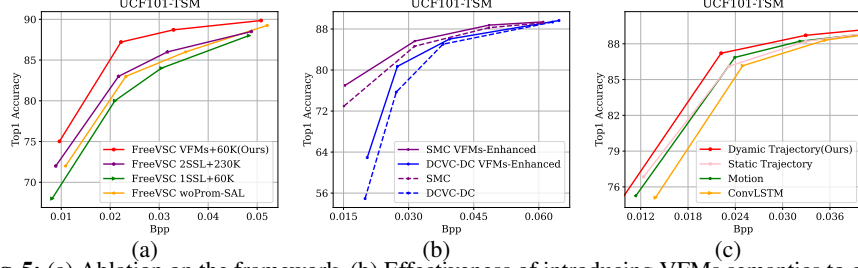


Fig. 5: (a) Ablation on the framework. (b) Effectiveness of introducing VFMs semantics to other approaches. (c) Comparison of different entropy models. ‘wo’ denotes without this component.

Effectiveness of Each Components. As depicted in Figure 5 (a), to demonstrate the superiority of rich semantics from VFMs, we first replace the VFMs-based objective with a self-supervised learning (SSL) objective, specifically the same MAE loss in SMC [89], training the model with the same 60K training videos as ours. The resulting FreeVSC 1SSL+60K model can be considered an enhanced version of SMC [89], incorporating our network architecture and employing the same learning objective as SMC, thus enabling a fair comparison of different semantic learning objectives. The FreeVSC 1SSL+60K model performs much worse than our model, exhibiting over a 5% accuracy drop at 0.02bpp. This proves that a single SSL objective learns much weaker semantic representations than pre-trained VFMs.

To investigate whether a more diverse SSL objective and training data are beneficial, we train the model with a combined objective [49] of MAE and contrastive learning, along with all 230K videos from Kinetics. The resulting FreeVSC 2SSL+230K model indeed improves performance over FreeVSC 1SSL+60K by approximately 2% due to learning richer semantics, but remains inferior to our model by about 2%. This underscores that previous unsupervised methods, such as SMC, are indeed constrained by their limited learning objective and data scale. In contrast, our approach directly absorbs and inherits the rich semantics from the free off-the-shelf VFMs, which are learned with more diverse objectives and more video data, resulting in stronger results.

Furthermore, we apply the proposed VFMs-based semantic learning scheme directly to DCVC-DC and SMC. As depicted in Figure 5 (b), we observe significant performance improvements in both methods, affirming the generalizability of our approach. The rich semantics from off-the-shelf VFMs proves consistently beneficial for addressing the semantic compression problem, regardless of the specific network architectures of the compression network.

Subsequently, we investigate the necessity of the Prom-SAL module. Concretely, we substitute the Prom-SAL module with a direct alignment scheme, employing simple linear layers to align the dimension of the compressed video semantics with that of VFM

features. The resulting FreeVSC woProm-SAL model exhibits a severe performance drop of 5% (see orange and red curves in Figure 5 (a)), attributed to the incapability of linear layers to effectively align the complex features from large-scale VFMs.

Finally, we demonstrate the superiority of our dynamic trajectory-based entropy coding scheme. We first replace the video-adaptive dynamic trajectory design with the static trajectory that is fixed and shared by all videos. As illustrated in Figure 5 (c), the performance degrades severely, since the fixed static trajectory can not well model the diverse contents of different videos. Then, we compare our approach to previous popular inter-frame coding scheme, namely the motion compensation-based [48] and ConvLSTM-based [109] schemes. It can be observed that both two models are inferior to ours, with approximately 1.5% and 2.2% accuracy drops for motion and ConvLSTM, respectively. The reasons may be that (1) explicit motion cannot effectively model high-level and abstract semantic dynamics, and (2) ConvLSTM struggles to adapt to video content. In contrast, our entropy model predicts customized non-linear transforms for each video, flexibly handling high-level information and adapting to video content.

Advantage of Learning from Multiple VFMs. Table 3 (a) presents the comparison of using different VFM. First, when employing a single VFM, the resulting models M_a (VideoMAEv2) and M_b (DINOv2) achieve unsatisfactory performances, with 83.11% and 83.56% Top1 accuracy, respectively. Furthermore, after leveraging both two above VFMs, the resulting model M_c attains better result 84.11% than using any single one of them, confirming that VFMs mutually enhance each other in our framework. Finally, scaling the number of VFMs by incorporating the SSL-Swin model [104], the final model M_{our} achieves the best result 85.07%.

Moreover, we visualize the feature maps of different models. The features are output from the third stage of the decoder network at the 0.02bpp level. As illustrated in Table 3 (b), the model trained with the SSL objective in SMC [89] exhibits weakest semantic representation capability, where the ‘person’ cannot be discriminated from the background effectively. In contrast, the model trained with VFM VideoMAEv2 demonstrates strong semantics in temporally moving regions, such as ‘person arm and leg’, while missing the static region ‘person head’. Further, the model equipped with 2VFMs VideoMAEv2+DINOv2 preserves semantics more comprehensively, demonstrating that our framework absorbs complementary semantics from diverse VFMs.

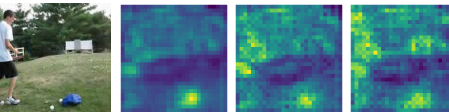
Ablation on Prom-SAL. As mentioned in Section 3, we introduce Prom-SAL, *i.e.*, a VFM-shared transformation ψ with VFM-specific prompts, to align the compressed video representation to features from different VFMs. In this section, we conduct extensive experiments to verify the effectiveness of each design in Prom-SAL.

As shown in Table 4 (a), after removing VFM-specific prompts, the resulting model M_d learns average features of different VFMs. The distorted average semantics leads

Table 3: (a) Comparison of using different VFMs. (b) Visualization of compressed video features.

<i>Models</i>	M_a	M_b	M_c	M_{our}
VideoMAEv2	✓		✓	✓
DINOv2		✓	✓	✓
SSL-Swin				✓
Accuracy (%)	83.11	83.56	84.11	85.07

(a)



Frame SSL VideoMAEv2 2VFMs

(b)

Table 4: (a) Ablation Study on Prom-SAL. C and F denote coarse and fine-level features, respectively. (b) *upper*: Comparison of different strategies for sharing the Prom-SAL parameters when aligning multiple VFMs. Gradient harmony is calculated as the cosine similarity between the gradients propagated from VideoMAEv2 and DINOv2. *lower*: Comparison of different network architecture choices for the alignment layer ψ in Prom-SAL.

<i>Models</i>	M_d	M_e	M_f	M_{our}	Parameter Share	None	Adapter	LORA	Prompts
Shared-Transformer	✓	✓	✓	✓	Accuracy (%)	83.01	84.23	84.97	85.07
Shared-Prompts	✓				Gradient Harmony	0.42	0.62	0.87	0.89
Feature Level	C+F	C-Only	F-Only	C+F	ψ architecture	3D CNN	ST-divided	Global	Vanilla
Accuracy (%)	80.92	83.07	84.28	85.07	Accuracy (%)	79.56	82.76		85.07

(a)

(b)

to a 4.15% accuracy drop. We then compare the impact of using different-level features. Using only coarse- or fine-level features yields unsatisfactory results, while our combination strategy achieves the best result, consistent with previous studies [65].

Further, we compare different strategies for sharing parameters of Prom-SAL across different VFMs. As depicted in Table 4 (b) *upper*, the absence of a parameter-sharing design results in a performance drop of over 2%, attributed to gradient contradictions among different VFMs, i.e., low gradient harmony. We then explore two recent strategies: Adapter [74] and LORA [45]. Adapter involves inserting VFM-specific ST-Adapter blocks [74] into the shared Transformer, while LORA utilizes a low-rank adaption technique [45], introducing VFM-specific rank-4 matrices to tune attention weights of the shared Transformer. ST-Adapter achieves the most inferior result, attributed to the inflexibility of the depth-3D convolution used. In contrast, more flexible methods like LORA and Prompts, initially proposed in the natural language processing (NLP) community, exhibit promising results, alongside high gradient harmony values. It is anticipated that leveraging more advanced parameter-sharing strategies from the NLP community will further enhance our framework.

Finally, we explore network architecture choices for Prom-SAL beyond the vanilla Transformer. As depicted in Table 4 (b) *lower*, employing 3D CNN or simplified spatial-temporal divided attentions [6] results in the accuracy drop of 5.51% and 2.31%, respectively. This proves the necessity of utilizing a flexible vanilla Transformer architecture for aligning the rich semantics from complex VFMs.

Ablation on Dynamic Trajectory-based Compression Scheme. As shown in Table 5 (a), after removing both the content-adaptive trajectory and region-wise trajectory designs, the resulting model M_g increases the bitcost by 21.7% compared to our model. Then, using a region-wise but static trajectory design leads to model M_h , which sees an 11.2% increase in bitcost. Conversely, employing a content-adaptive yet region-shared trajectory design leads to model M_i , causing a 13.8% increase in bitcost. Additionally, comparing models M_j and M_{our} reveals the benefits of long-term information in semantic compression. Furthermore, investigating the impact of trajectory number S , we observe BD-rates of 21.7%, 6.4%, 0%, and -1.8% for $S = 1, 3, 5$, and 10, respectively. We select $S = 5$ to strike the best computational cost-performance trade-off.

Further, we visualize the predicted future frame features. As shown in Table 5 (b), the predicted future frame (the third figure), obtained from the two past historical frame

Table 5: (a) Ablation Study on Entropy Model. BD-Rate indicates the increased bitrates compared to our model. *Lower is better.* (b) Visualization of the predicted future semantic feature.

	M_g	M_h	M_i	M_j	M_{our}
Content-adaptive			✓	✓	✓
Region-wise		✓		✓	✓
Long-term Fusion	✓	✓	✓		✓
BD-Rate (%)	21.7	11.2	13.8	7.6	0

(a)

Historical two frames Predicted Ground-truth

(b)

features (the first two figures), accurately reflects the future semantics ‘the man bent down after throwing’ of the ground-truth frame (the fourth figure).

Model Complexity. The parameter numbers of our encoder and decoder networks are 10.4M and 9.3M, respectively. We report the per frame running time of a 1080p video on the machine with a Nvidia 2080Ti GPU. Our encoding time is 722ms, which is better than DCVC-DC (1005ms) and SMC (1413ms). Our decoding time is 226ms, which is $3\times$ faster than DCVC-DC (765ms). While our model is trained with several large VFMs, they are only utilized during training. Therefore, our model still maintains a fast inference speed, since most layers of our networks are operated on low-resolution feature maps ($\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ resolution scales).

5 Conclusion, Limitation, and Potential Negative Impact

Conclusion. In this paper, we have introduced Free-VSC, the first video semantic compression framework leveraging VFMs. To enable VFMs to collaboratively guide the learning of the compression model, a prompt-based semantic alignment layer is introduced. Additionally, we have proposed a trajectory-based dynamic entropy model for more efficient inter-frame semantic compression. Comprehensive experiments validate our approach’s superiority in three video analysis tasks.

Limitation and Future Work. First, the number of VFMs is rapidly increasing. It is possible that some of them are not helpful to the semantic compression task. A fast method for judging their effectiveness is missed in this work and will be developed in the future. Second, same VFMs are employed across different bitrates in this work for simplicity. However, different VFMs usually contain knowledge of different granularity. A bitrate-adaptive VFM selection strategy will be devised in the future. Besides, some hyper-parameters in this work rely on empirical selection, which can be automatically decided by the reinforce learning algorithms.

Potential Negative Impact. The proposed Free-VSC extends the boundaries of the minimum bitcost required to preserve general semantic information within videos. The concept and technique are general and unbiased toward any specific video task, as we avoid using task-specific annotations. However, despite its technical contributions and our original intention not to favor any particular application, concerns remain regarding its potential illegal deployment in applications that may violate privacy rights and human rights, such as video surveillance and behavior monitoring.

Acknowledgment. This work is supported by National Natural Science Foundation of China (62225112), Shanghai Artificial Intelligence Laboratory, and National Natural Science Foundation of China under Grant 62101326.

References

1. Mmtracking: Openmmlab video perception toolbox and benchmark. <https://github.com/open-mmlab/mmtracking> (2020)
2. Openmmlab's next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmaction2> (2020)
3. Akbari, M., Liang, J., Han, J.: Dsslic: Deep semantic segmentation-based layered image compression. In: International Conference on Acoustics, Speech and Signal Processing (2019)
4. Bai, Y., Yang, X., Liu, X., Jiang, J., Wang, Y., Ji, X., Gao, W.: Towards end-to-end image compression and analysis with transformers. In: Annual AAAI Conference on Artificial Intelligence (2022)
5. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (2018)
6. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: International Conference on Machine Learning (2021)
7. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology* (2021)
8. Cai, Q., Chen, Z., Wu, D.O., Liu, S., Li, X.: A novel video coding strategy in hevc for object detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2021)
9. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in Neural Information Processing Systems* (2020)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
12. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
13. Chao, J., Steinbach, E.: Keypoint encoding for improved feature extraction from compressed video at low bitrates. *IEEE Transactions on Multimedia* (2015)
14. Che, Z., Borji, A., Zhai, G., Ling, S., Li, J., Tian, Y., Guo, G., Le Callet, P.: Adversarial attack against deep saliency models powered by non-redundant priors. *IEEE Transactions on Image Processing* (2021)
15. Chen, H., Qu, Z., Tian, Y., Jiang, N., Qin, Y., Gao, J., Zhang, R., Ma, Y., Jin, Z., Zhai, G.: A cross-temporal multimodal fusion system based on deep learning for orthodontic monitoring. *Computers in Biology and Medicine* (2024)
16. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. *arXiv* (2020)
17. Chen, Y.H., Weng, Y.C., Kao, C.H., Chien, C., Chiu, W.C., Peng, W.H.: Transtic: Transferring transformer-based image compression from human perception to machine perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
18. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

19. Chen, Z., Fan, K., Wang, S., Duan, L.Y., Lin, W., Kot, A.: Lossy intermediate deep learning feature compression and evaluation. In: ACM International Conference on Multimedia (2019)
20. Chen, Z., Fan, K., Wang, S., Duan, L., Lin, W., Kot, A.C.: Toward intelligent sensing: Intermediate deep feature compression. *IEEE Transactions on Image Processing* (2019)
21. Chen, Z., Sun, W., Tian, Y., Jia, J., Zhang, Z., Wang, J., Huang, R., Min, X., Zhai, G., Zhang, W.: Gaia: Rethinking action quality assessment for ai-generated videos. *arXiv* (2024)
22. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv* (2022)
23. Choi, H., Bajic, I.V.: High efficiency compression for object detection. In: International Conference on Acoustics, Speech and Signal Processing (2018)
24. Choi, H., Bajić, I.V.: Near-lossless deep feature compression for collaborative intelligence. In: International Workshop on Multimedia Signal Processing (2018)
25. Choi, H., Bajic, I.V.: Scalable image coding for humans and machines. *IEEE Transactions on Image Processing* (2022)
26. Choi, J., Han, B.: Task-aware quantization network for jpeg image compression. In: European Conference on Computer Vision (2020)
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
28. Duan, H., Shen, W., Min, X., Tian, Y., Jung, J.H., Yang, X., Zhai, G.: Develop then rival: A human vision-inspired framework for superimposed image decomposition. *IEEE Transactions on Multimedia* (2022)
29. Duan, L.Y., Chandrasekhar, V., Chen, J., Lin, J., Wang, Z., Huang, T., Girod, B., Gao, W.: Overview of the mpeg-cdvs standard. *IEEE Transactions on Image Processing* (2015)
30. Duan, L.Y., Gao, F., Chen, J., Lin, J., Huang, T.: Compact descriptors for mobile visual search and mpeg cdvs standardization. In: IEEE International Symposium on Circuits and Systems (2013)
31. Duan, L.Y., Lou, Y., Bai, Y., Huang, T., Gao, W., Chandrasekhar, V., Lin, J., Wang, S., Kot, A.C.: Compact descriptors for video analysis: The emerging mpeg standard. *IEEE Transactions on Multimedia* (2018)
32. Duan, L., Liu, J., Yang, W., Huang, T., Gao, W.: Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *IEEE Transactions on Image Processing* (2020)
33. Duan, S., Chen, H., Gu, J.: Jpd-se: High-level semantics for joint perception-distortion enhancement in image compression. *IEEE Transactions on Image Processing* (2022)
34. Dubois, Y., Bloem-Reddy, B., Ullrich, K., Maddison, C.J.: Lossy compression for lossless prediction. In: Advances in Neural Information Processing Systems (2021)
35. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
36. Fang, Z., Shen, L., Li, M., Wang, Z., Jin, Y.: Prior-guided contrastive image compression for underwater machine vision. *IEEE Transactions on Circuits and Systems for Video Technology* (2022)
37. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
38. Feng, R., Jin, X., Guo, Z., Feng, R., Gao, Y., He, T., Zhang, Z., Sun, S., Chen, Z.: Image coding for machines with omnipotent feature learning. *arXiv* (2022)
39. Galteri, L., Bertini, M., Seidenari, L., Del Bimbo, A.: Video compression for object detection algorithms. In: International Conference on Pattern Recognition (2018)

40. Ge, X., Luo, J., Zhang, X., Xu, T., Lu, G., He, D., Geng, J., Wang, Y., Zhang, J., Qin, H.: Task-aware encoder control for deep video compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
41. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* (2020)
42. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning (2020)
43. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
44. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
45. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. *arXiv* (2021)
46. Hu, Y., Yang, S., Yang, W., Duan, L.Y., Liu, J.: Towards coding for human and machine vision: A scalable image coding approach. In: International Conference on Multimedia and Expo (2020)
47. Hu, Z., Lu, G., Guo, J., Liu, S., Jiang, W., Xu, D.: Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
48. Hu, Z., Lu, G., Xu, D.: Fvc: A new framework towards deep video compression in feature space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
49. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
50. Huang, Z., Jia, C., Wang, S., Ma, S.: Visual analysis motivated rate-distortion model for image coding. In: International Conference on Multimedia and Expo (2021)
51. Huang, Z., Jia, C., Wang, S., Ma, S.: Hmfvc: A human-machine friendly video compression scheme. *IEEE Transactions on Circuits and Systems for Video Technology* (2022)
52. Huynh-Thu, Q., Garcia, M.N., Speranza, F., Corriveau, P., Raake, A.: Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting* (2010)
53. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
54. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (2022)
55. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008)
56. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv* (2014)
57. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2011)
58. Li, J., Li, B., Lu, Y.: Deep contextual video compression. In: Advances in Neural Information Processing Systems (2021)

59. Li, J., Li, B., Lu, Y.: Neural video compression with diverse contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
60. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
61. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
62. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: European Conference on Computer Vision (2018)
63. Lin, H., Chen, B., Zhang, Z., Lin, J., Wang, X., Zhao, T.: Deepsvic: Deep scalable video coding for both machine and human vision. In: ACM International Conference on Multimedia (2023)
64. Lin, J., Gan, C., Han, S.: Tsm: temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
65. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
66. Liu, J., Wang, S., Ma, W.C., Shah, M., Hu, R., Dhawan, P., Urtasun, R.: Conditional entropy coding for efficient video compression. In: European Conference on Computer Vision (2020)
67. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
68. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: Dvc: An end-to-end deep video compression framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
69. Lu, G., Zhang, X., Ouyang, W., Chen, L., Gao, Z., Xu, D.: An end-to-end learning framework for video compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020)
70. Mentzer, F., Toderici, G.D., Minnen, D., Caelles, S., Hwang, S.J., Lucic, M., Agustsson, E.: Vct: A video compression transformer. In: Advances in Neural Information Processing Systems (2022)
71. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. *arXiv* (2016)
72. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: IEEE International Conference on Image Processing (2020)
73. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv* (2023)
74. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. In: Advances in Neural Information Processing Systems (2022)
75. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (2019)
76. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. *arXiv* (2017)
77. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

78. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal (1948)
79. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
80. Singh, S., Abu-El-Haija, S., Johnston, N., Ballé, J., Shrivastava, A., Toderici, G.: End-to-end learning of compressible features. In: IEEE International Conference on Image Processing (2020)
81. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv (2012)
82. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on Circuits and Systems for Video Technology (2012)
83. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N.: Diverse semantic image synthesis via probability distribution modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
84. Tian, Y., Che, Z., Bao, W., Zhai, G., Gao, Z.: Self-supervised motion representation via scattering local motion cues. In: European Conference on Computer Vision (2020)
85. Tian, Y., Che, Z., Zhai, G., Gao, Z.: Ban, a barcode accurate detection network. In: IEEE Visual Communications and Image Processing (2018)
86. Tian, Y., Lu, G., Min, X., Che, Z., Zhai, G., Guo, G., Gao, Z.: Self-conditioned probabilistic learning of video rescaling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
87. Tian, Y., Lu, G., Yan, Y., Zhai, G., Chen, L., Gao, Z.: A coding framework and benchmark towards low-bitrate video understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
88. Tian, Y., Lu, G., Zhai, G.: Smc++: Masked learning of unsupervised video semantic compression. arXiv (2024)
89. Tian, Y., Lu, G., Zhai, G., Gao, Z.: Non-semantics suppressed mask learning for unsupervised video semantic compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
90. Tian, Y., Min, X., Zhai, G., Gao, Z.: Video-based early asd detection via temporal pyramid networks. In: International Conference on Multimedia and Expo (2019)
91. Tian, Y., Yan, Y., Zhai, G., Chen, L., Gao, Z.: Clsa: a contrastive learning framework with selective aggregation for video rescaling. IEEE Transactions on Image Processing (2023)
92. Tian, Y., Yan, Y., Zhai, G., Guo, G., Gao, Z.: Ean: event adaptive network for enhanced action recognition. International Journal of Computer Vision (2022)
93. Tomar, S.: Converting video formats with ffmpeg. Linux Journal (2006)
94. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022)
95. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
96. Veselov, A.I., Chen, H., Romano, F., Zhijie, Z., Gilmudtinov, M.R.: Hybrid video and feature coding and decoding (2021), uS Patent App. 17/197,500
97. Wang, J., Chen, D., Wu, Z., Luo, C., Tang, C., Dai, X., Zhao, Y., Xie, Y., Yuan, L., Jiang, Y.G.: Look before you match: Instance understanding matters in video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)

98. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
99. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
100. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv (2022)
101. Wieckowski, A., Brandenburg, J., Hinz, T., Bartnik, C., George, V., Hege, G., Helmrich, C., Henkel, A., Lehmann, C., Stoffers, C., Zupancic, I., Bross, B., Marpe, D.: Vvenc: An open and optimized vvc encoder implementation. In: IEEE International Conference on Multimedia and Expo Workshops
102. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the h. 264/avc video coding standard. IEEE Transactions on Circuits and Systems for Video Technology (2003)
103. Wu, C.Y., Singhal, N., Krahenbuhl, P.: Video compression through image interpolation. In: European Conference on Computer Vision (2018)
104. Xie, Z., Lin, Y., Yao, Z., Zhang, Z., Dai, Q., Cao, Y., Hu, H.: Self-supervised learning with swin transformers. arXiv (2021)
105. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv (2015)
106. Yan, Z., Li, G., Tian, Y., Wu, J., Li, S., Chen, M., Poor, H.V.: Dehib: Deep hidden backdoor attack on semi-supervised learning via adversarial perturbation. In: Annual AAAI Conference on Artificial Intelligence (2021)
107. Yan, Z., Li, S., Zhao, R., Tian, Y., Zhao, Y.: Dhbe: Data-free holistic backdoor erasing in deep neural networks via restricted adversarial distillation. In: ACM ASIA Conference on Computer and Communications Security (2023)
108. Yang, F., Herranz, L., Van De Weijer, J., Guitián, J.A.I., López, A.M., Mozerov, M.G.: Variable rate deep image compression with modulated autoencoder. IEEE Signal Processing Letters (2020)
109. Yang, R., Mentzer, F., Van Gool, L., Timofte, R.: Learning for video compression with recurrent auto-encoder and recurrent probability model. IEEE Journal of Selected Topics in Signal Processing (2020)
110. Yang, R., Timofte, R., Van Gool, L.: Advancing learned video compression with in-loop frame prediction. IEEE Transactions on Circuits and Systems for Video Technology (2022)
111. Yang, R., Van Gool, L., Timofte, R.: Perceptual learned video compression with recurrent conditional gan. arXiv (2021)
112. Yang, Z., Wang, Y., Xu, C., Du, P., Xu, C., Xu, C., Tian, Q.: Discernible image compression. In: ACM International Conference on Multimedia (2020)
113. Yi, C., Yang, S., Li, H., Tan, Y.p., Kot, A.: Benchmarking the robustness of spatial-temporal models against corruptions. In: Advances in Neural Information Processing Systems (2021)
114. Yi, F., Chen, M., Sun, W., Min, X., Tian, Y., Zhai, G.: Attention based network for no-reference ugc video quality assessment. In: IEEE International Conference on Image Processing (2021)
115. Zhang, F., Bull, D.R.: A parametric framework for video compression using region-based texture models. IEEE Journal of Selected Topics in Signal Processing (2011)
116. Zhang, Q., Wang, S., Zhang, X., Ma, S., Gao, W.: Just recognizable distortion for machine vision oriented image and video coding. International Journal of Computer Vision (2021)
117. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

118. Zhang, X., Wu, X.: Lvqac: Lattice vector quantization coupled with spatially adaptive companding for efficient learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
119. Zhang, X., Ma, S., Wang, S., Zhang, X., Sun, H., Gao, W.: A joint compression scheme of video feature descriptors and visual content. *IEEE Transactions on Image Processing* (2016)
120. Zhang, X., Ge, X., Xu, T., He, D., Wang, Y., Qin, H., Lu, G., Geng, J., Zhang, J.: Gaussianimage: 1000 fps image representation and compression by 2d gaussian splatting. *arXiv* (2024)
121. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Byte-track: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision (2022)
122. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. In: International Conference on Machine Learning (2024)
123. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: International Conference on Learning Representations (2021)