V1-Inspired Dynamic Vision System: A Bio-Plausible Video Embedding Framework with Decoupled Shape–Color Pathways and Long-Range Spatiotemporal Perception

Beier Meng
Department of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
2022211920@bupt.cn

Abstract

Large-scale video–text contrastive learning has become the de facto standard for video understanding, yet mainstream methods almost universally adopt the linear patch embedding of vision Transformers (ViT) or evenly sliced tubelets as the front end. We revisit this front end from the biological principles of the primary visual cortex (area V1) and introduce **V1DynamicVisionSystem (V1-DVS)**, a fully end-to-end trainable, bio-plausible video embedding framework. Its main contributions include (i) **three processing components in two parallel streams**, with serial layer 4C orientation and layer 4B motion-direction processing alongside the cytochrome-oxidase Blob color-opponent stream, (ii) **explicit shape–color decoupling** with isotropic Blob processing and anisotropic Gabor orientation processing, (iii) **long-range spatiotemporal convolution** that encodes inter-frame motion through 3D filters and the Adelson–Bergen motion-energy model rather than naive frame stacking, and (iv) **bidirectional cross-pathway attention**, allowing motion and color features to mutually modulate, as in V1$\rightarrow$V4 inter-areal feedback. Trained on CC3M, WebVid-10M, ActivityNet and Ego4D with self-supervised objectives, V1-DVS achieves higher self-R@1 than four reference front ends on ActivityNet and Ego4D, and leads on four of five ActivityNet embedding-geometry metrics under matched from-scratch training.

1 Introduction and Background

Mainstream video–text embedding models inherit their front end almost unchanged from the image domain: a single content-agnostic linear projection maps each patch—or each evenly sliced tubelet—into a token [14,15]. Oriented form, coherent motion, and chromatic organization are then learned implicitly. An edge prior can encourage shape-sensitive features [26]; explicit motion processing can counter the bag-of-frames behaviour observed in video–language learning [23]; and embedding geometry provides diagnostics complementary to retrieval [46]. A principled front end should therefore supply structure that video shares with natural visual input.

The mammalian visual system compresses a high-dimensional spatiotemporal stream into a compact, behaviourally useful code through factorized decompositions organized into parallel pathways. This motivates our central hypothesis: oriented form, phase-invariant motion, and color opponency, transcribed into differentiable operators, may improve temporal fidelity and embedding geometry. We test this hypothesis through matched front-end comparisons and component ablations. We therefore review the primary visual cortex (V1, Brodmann area 17) [1,2] as a template for computational roles in our front end.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1.1 Biological Background

The neural substrate of human visual perception is the primary visual cortex (V1, Brodmann area 17) [1,2]. Hubel and Wiesel's classical work revealed that V1 is organized perpendicular to its surface into orientation, ocular-dominance and spatial-frequency columns, which jointly form the smallest functional unit of visual processing—the *hypercolumn*—with neighbouring orientation columns spanning $\sim 180^\circ$ smoothly, so that any local field is covered by a complete set of tunings [3]. V1 is laminated into layers 1–6, with layer 4C as the principal recipient of LGN input: 4Cα receives magnocellular drive (high-temporal/low-spatial-frequency, motion sensitive), while 4Cβ receives parvocellular drive (fine spatial detail and color) [4]. Simple cells in 4C have spatially separated ON/OFF subregions accurately modeled by Gabor functions [5,6]; orientation selectivity arises from in-phase arrangement of LGN inputs together with intracortical lateral inhibition.

Layer 4B introduces motion direction selectivity: unlike orientation-only 4C cells, 4B cells distinguish 0° from 180° via Barlow–Levick asymmetric inhibition or Reichardt-style delay-line detection [7]. Adelson and Bergen formalized this as a quadrature pair of spatiotemporal filters whose squared sum yields a phase-invariant directional response, $E_{\mathrm{dir}} = F_{\mathrm{even}}^2 + F_{\mathrm{odd}}^2$ [8]. Hypercomplex cells in 4B further exhibit *end-stopping*—an excitatory center with inhibitory surround—and respond strongly to endpoints and corners but weakly to long-line interiors, providing a substrate for closure and curvature [9]. Parallel to this form/motion stream, V1 hosts a dedicated color stream: Blob regions in layer 2/3 contain double-opponent cells with both spatial and chromatic antagonism (e.g. center R+/G−, surround R−/G+), making them sensitive to chromatic edges and supporting color constancy [10,11]. Livingstone and Hubel showed that V1 thus carries two anatomically parallel pathways—inter-blob (form) and blob (color)—establishing the neural basis for separated form and color representations [12].

1.2 Image and Video Embedding Models

In the deep-learning era, visual embedding has transitioned from convolutional networks to Transformers. CNN architectures such as ResNet [13] obtain hierarchical features but their kernels are randomly initialized and require huge data to implicitly learn Gabor-like orientation tuning. Vision Transformers (ViT) [14] split images into 16×16 patches with a linear projection, achieving breakthrough results but discarding the convolutional local prior altogether. ViViT [15], TimeSformer [16] and VideoMAE [17] extend ViT to video—ViViT proposes evenly sliced tubelet embeddings, TimeSformer factorizes spatial and temporal attention, and VideoMAE leverages 75% masked-tube reconstruction. These models are typically pretrained with large-scale video–text contrastive learning; CLIP [18] aligned vision and language on 400M image–text pairs via InfoNCE, and CLIP4Clip [19], VideoCLIP [20], InternVideo [21] extended this to video. While efficient, the pixel-to-patch linear projection treats the input uniformly. Our central question is therefore: *can a front-end inspired by V1's high-level decompositions endow video embeddings with stronger temporal fidelity and better isotropy, under matched from-scratch training?*

2 Related Work

Video semantic embedding. Video–text contrastive pretraining (VideoCLIP [20], CLIP4Clip [19], InternVideo [21]) achieves strong generalization on huge corpora; later works such as Frozen in Time [22], Singularity [23], TimeSformer [16] and VideoMAE [17] explore alternative temporal aggregations or masked self-supervised objectives. Tubelet embeddings (ViViT [15], MViT [24]) account for the time axis but their 3D kernels are still randomly initialized and offer no inductive bias for motion.

Bio-inspired vision networks. VOneNet [25] inserts a fixed Gabor front end before a classification network, improving adversarial robustness and out-of-domain generalization. BagNet [26] models the texture bias of CNNs through small local receptive fields, exposing the over-reliance of CNNs on texture rather than shape. CORnet [27] maps the ventral stream onto deep network layers. In contrast to these works, we model the multi-pathway structure of V1 as a fully end-to-end trainable video embedding module and, for the first time in this setting, introduce bidirectional cross-pathway attention together with a multi-objective self-supervised pretraining recipe.

Self-supervised pretraining. CLIP [18] established the vision–language contrastive paradigm; BEiT [28] and MAE [29] introduced masked image/token prediction for stronger patch-level features. SimCLR [41] showed that strong augmentations and an NT-Xent objective alone yield competitive image representations, while Misra et al. [42] introduced temporal-order verification (*shuffle and learn*) as a pretext task for video. Our self-supervised recipe combines these three families (NT-Xent, temporal order, masked token prediction) into a CLIP-free objective tailored to the bio-inspired video encoder.

3 Method

V1-DVS (Figure 1) follows a four-stage design: *two-stream processing* → *cross-pathway fusion* → *dynamic patching* → *semantic projection*. An input video

$$x \in \mathbb{R}^{B\times 3\times T\times H\times W} \tag{1}$$

is first dispatched into two parallel streams: (i) a **shape–motion** pathway (4C→4B in series) that extracts orientation selectivity and motion-direction energy, and (ii) a **color** pathway (Blob) that extracts color opponency and chromatic edges. The two streams are merged at a unified spatial resolution by a bidirectional cross-attention module, transformed into a token sequence by a V1 dynamic patch embedding, and finally projected to the semantic space by a shallow Transformer encoder. To balance biological fidelity with the computational demands of large-scale training, every pathway is implemented through systematic optimizations—shared-basis Gabor projection, low-rank spatiotemporal factorization, and a depthwise-separable color stream—that preserve all biological abstractions while reducing parameters and memory. We describe each module below.

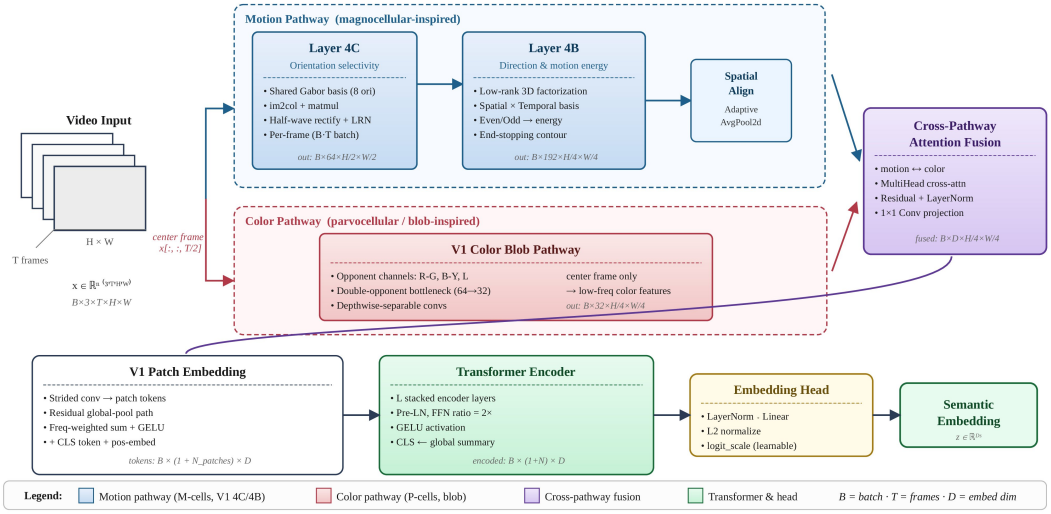


Figure 1: Overall architecture of V1-DVS. The shape–motion stream applies 4C orientation processing followed by 4B motion-energy processing. Bidirectional cross-pathway attention fuses this stream with the parallel Blob color stream before dynamic patch embedding and shallow Transformer projection.

3.1 V1C Pathway: 4C-Layer Orientation Selectivity

A single orientation column responds to one orientation only; the strength of V1 lies in many such columns operating in parallel. A naive implementation builds a separate 2D convolution for every orientation, which is parameter-redundant and triggers N_{ori} separate cuDNN launches. We design a **shared-basis projection with intra-column mixing** (Figure 2), fusing all column responses into a single matrix multiplication. Given an input feature map $x \in \mathbb{R}^{B\times C\times H\times W}$ and a shared Gabor basis $\mathbf{B} \in \mathbb{R}^{N_b\times K^2}$, we unfold x via im2col to obtain $\mathbf{U} \in \mathbb{R}^{B\times C\times K^2\times L}$, then perform the basis

projection

$$\mathbf{Z}_{b,c,n,\ell} = \sum_{k=1}^{K^2} \mathbf{B}_{n,k}\, \mathbf{U}_{b,c,k,\ell}, \qquad \mathbf{Z}_{b,c} = \mathbf{B}\, \mathbf{U}_{b,c} \in \mathbb{R}^{N_b \times L}. \tag{2}$$

Stacking the slices gives $\mathbf{Z} \in \mathbb{R}^{B\times C\times N_b\times L}$. Merging the (C, N_b) dimensions yields $\tilde{\mathbf{Z}} \in \mathbb{R}^{B\times (CN_b)\times L}$. An **inter-column mixing** step then linearly recombines the basis responses into the final column outputs through a learnable matrix $\mathbf{W}_{\text{mix}} \in \mathbb{R}^{N_o\times (CN_b)}$ (with $N_o = N_{\text{ori}} \cdot C_{\text{per}}$):

$$\mathbf{Y}_{b,o,\ell} = \sum_{j=1}^{CN_b} \mathbf{W}_{\text{mix},o,j}\, \tilde{\mathbf{Z}}_{b,j,\ell}, \qquad \mathbf{Y} \in \mathbb{R}^{B\times N_o\times L}. \tag{3}$$

$\mathbf{W}_{\text{mix}}$ is initialized so that the C_{per} channels under a column bind to the corresponding orientation index of the basis. This realises the *population response* of an orientation hypercolumn—the joint activity over all columns encodes local orientation, decodable by a population vector [30]. The output is reshaped to $[B, N_o, H, W]$ and processed by three biologically motivated stages: (a) **half-wave rectification** for unilateral simple-cell activation; (b) **cross-orientation local normalization** for V1 surround suppression [31]; and (c) **intra-column grouped mixing** via a 1×1 grouped convolution with N_{ori} groups, which allows same-column interactions while preventing cross-column leakage. A final max-pool aggregates simple→complex cells, yielding the spatial component of the Adelson–Bergen energy model [8].

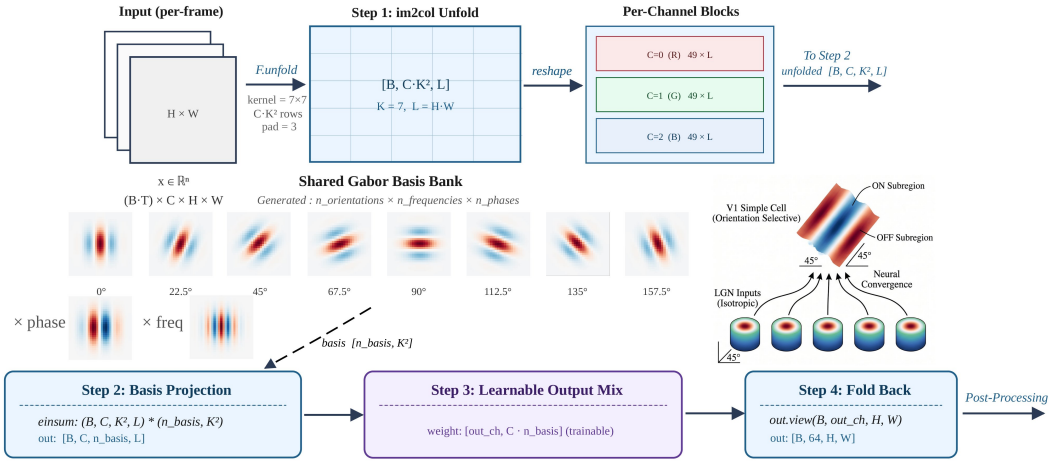


Figure 2: The V1C pathway. A shared Gabor basis is mixed into orientation columns and post-processed by rectification, surround normalization and grouped mixing.

The core of the V1C pathway is an orientation-selective Gabor filter bank. A Gabor function—the product of a Gaussian envelope and a sinusoidal carrier—minimizes joint uncertainty in space, frequency and orientation [5] and matches cat V1 simple cells empirically [6]. For each (θ, f, ϕ), we generate a kernel $G_{\theta,f,\phi}(x, y) = \exp\big(-(x_{\text{rot}}^2 + y_{\text{rot}}^2)/(2\sigma^2)\big)\cos\big(2\pi f x_{\text{rot}}/K + \phi\big)$. Here K is the kernel width in samples, f is the nominal frequency across that width, and $\phi \in \{0, \pi/2\}$ gives the even and odd phases. The default N_{ori}=8 uses 22.5° steps inspired by cortical organization [3]. We mean-subtract and L_2-normalize each kernel. Because V1 simple cells are tuned in spatial frequency as well, with preferred frequencies spanning ∼0.5–8 cpd [4], we use N_{freq}=3 nominal frequencies over a low band of approximately 0.5–2, a computational subset constrained by the K=7 kernel support. We enumerate all triples (θ, f, ϕ) to form a shared basis bank $\mathbf{B} \in \mathbb{R}^{N_b\times K^2}$ with $N_b = N_{\text{ori}}\, N_{\text{freq}}\, N_{\text{phase}}$. This bank is shared across columns, replacing N_{ori} separate Gabor convolutions with a small structured (and optionally learnable) basis, providing a structured prior that yields meaningful local edge responses from step zero and improves low-data generalization [25].

3.2 V1B Pathway: 4B-Layer Motion Direction Selectivity

3.2.1 Low-Rank Spatiotemporal Factorization for Direction Selectivity

The V1B pathway (Figure 3) takes the V1C feature sequence $[B, C_{\text{in}}, T, H, W]$ and extracts motion-direction features—classically a 3D spatiotemporal filter selecting oblique space–time energy. A naive implementation that builds an independent 3D conv for every direction and phase consumes >2M parameters at typical sizes and becomes the parameter bottleneck of the system. We design a **low-rank spatiotemporal factorization** that decomposes the 3D conv into four cheap operators applied in sequence:

$$\text{Conv3d}(C_{\text{in}}, C_{\text{out}}, T_k, K, K) \;\approx\; \phi_{\text{out}} \circ \phi_{\text{temporal}} \circ \phi_{\text{spatial}} \circ \phi_{\text{in}}, \tag{4}$$

where ϕ_{in} is a $1\times1\times1$ conv compressing channels to a low rank R, ϕ_{spatial} is a depthwise spatial basis shared across directions, ϕ_{temporal} is a shared depthwise temporal basis ($T_k\times1\times1$ conv) modeling temporal dynamics, and ϕ_{out} is a per-direction readout matrix $\mathbf{M}^{(d)} \in \mathbb{R}^{C_{\text{per}}\times R}$. The output for direction d is

$$\mathbf{F}^{(d)}_{b,c,h,w} = \sum_{r=1}^{R} \mathbf{M}^{(d)}_{c,r}\, \mathbf{z}_{b,r,t_*,h,w}, \qquad t_* = \lfloor T/2 \rfloor. \tag{5}$$

This factorization brings the cost down to roughly 20% of the naive implementation, and biologically corresponds to "front-end channel compression $\rightarrow$ shared spatial RFs $\rightarrow$ shared temporal filtering $\rightarrow$ direction-specific readout" across V1 layers. The low-rank space R acts as a set of spatiotemporal basis functions shared by all directions, and direction differences live entirely in the linear readout, matching the empirical finding that V1 cells share substrate RFs and encode direction by phase combinations [30]. We initialize the readout $\mathbf{M}^{(d)}$ with cosine-phase encoding along the rank dimension so that neighbouring directions vary smoothly and the phase-symmetry of Reichardt delay-line detectors [7,32] is satisfied. A **depthwise-separable receptive-field expansion** module ("5×5 depthwise + 1×1 pointwise + dilated 3×3 depthwise + 1×1 pointwise") is applied at the V1B front, giving an effective RF roughly 1.5–2.5$\times$ that of V1C, in line with biology [9].

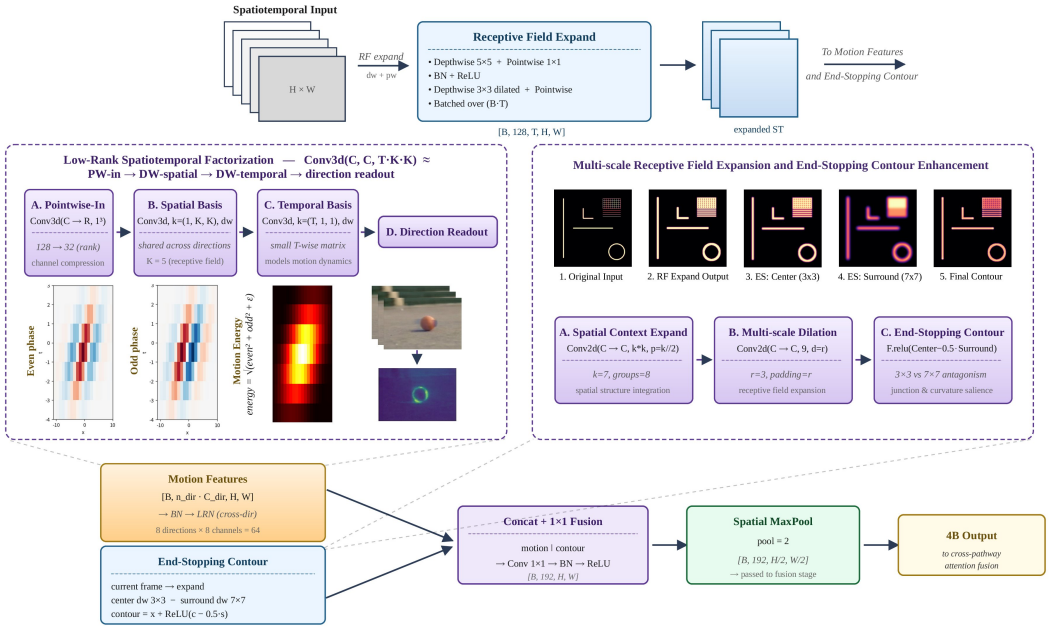


Figure 3: The V1B pathway. Low-rank factorization with quadrature even/odd readouts implements the Adelson–Bergen motion-energy model; end-stopping enhances corners.

3.2.2 Adelson–Bergen Motion-Energy Model

The motion-energy model [8] formalizes phase-invariant direction responses as the sum of squares of a quadrature pair, equivalent to projecting the spatiotemporal Fourier energy onto a chosen direction;

this matches the behaviour of complex cells. We embed this model into the low-rank framework through **a shared low-rank basis with quadrature readouts**: the even and odd direction readouts *share* the front ϕ_{in}, spatial basis ϕ_{spatial} and temporal basis ϕ_{temporal}, but maintain two independent direction matrices $\mathbf{M}^{(d)}_{\text{even}}, \mathbf{M}^{(d)}_{\text{odd}}$. Given the shared intermediate $\mathbf{z}_{t_*}$,

$$\mathbf{F}^{(d)}_{\text{even},c,h,w} = \sum_r \mathbf{M}^{(d)}_{\text{even},c,r}\, \mathbf{z}_{t_*,r,h,w}, \quad \mathbf{F}^{(d)}_{\text{odd},c,h,w} = \sum_r \mathbf{M}^{(d)}_{\text{odd},c,r}\, \mathbf{z}_{t_*,r,h,w}, \tag{6}$$

with the two readouts initialized as $\cos(\varphi_r + \phi_d)$ and $\sin(\varphi_r + \phi_d)$ so that they form an orthogonal Hilbert pair in rank space. The motion energy follows as

$$E^{(d)}_{c,h,w} = \sqrt{(\mathbf{F}^{(d)}_{\text{even},c,h,w})^2 + (\mathbf{F}^{(d)}_{\text{odd},c,h,w})^2 + \epsilon}. \tag{7}$$

This enjoys phase invariance, non-negativity (no extra ReLU needed), and a magnitude response whose contrast dependence is controlled by normalization. Combined with cross-direction local normalization, it yields a winner-take-all competition over directions, producing sparse and salient direction codes. By sharing the first three operators, this design further reduces the cost of motion-energy computation by about 40% relative to a naive even/odd convolution pair, without sacrificing phase-coding capacity.

3.2.3 End-Stopping and Contour Enhancement

Some 4B cells exhibit end-stopping (hypercomplex cells), responding weakly to the middle of long lines and strongly to corners and endpoints [9,33], via center–surround antagonism: an excitatory short-segment center suppressed by a continuation into the surround. Our **dual-scale end-stopping** module uses a small-kernel (3×3) depthwise conv for the center and a large-kernel (7×7) depthwise conv for the surround, adding back the contour signal as a residual,

$$F_{\text{contour}} = F + \max\big(F_{\text{center}} - 0.5\, F_{\text{surround}},\, 0\big). \tag{8}$$

This produces strong responses at corners, T-junctions and curvature discontinuities while being near zero in the middle of long lines, sparsifying the feature map. Depthwise convolutions compress parameters by a factor of $1/C$ relative to dense convs, so end-stopping comes at near-zero parameter cost. The V1B output concatenates E and F_{contour} along channels, fuses them by a 1×1 conv, and downsamples spatially.

3.3 Color Pathway: Blob/Inter-Blob System

3.3.1 Color-Opponent Matrix and Center–Surround Antagonism

The first stage is a **fixed color-opponent projection** (Figure 4): the input $I \in \mathbb{R}^{B\times3\times H\times W}$ is mapped through a 1×1 conv whose weights are fixed at biologically meaningful color opponents (approximating the DKL space [34]), $I_{\text{opp}} = W_{\text{opp}}\, I \in \mathbb{R}^{B\times6\times H\times W}$. The six rows represent L$-$M, M$-$L, S$-$(L+M), (L+M)$-$S, luminance ON, and luminance OFF, mirroring the parvocellular, koniocellular and magnocellular pathways of the retina/LGN [35,36]. The transform leaves the highly correlated RGB cone basis for a decorrelated, physiologically meaningful color coordinate system, naturally amplifying chromatic contrasts such as "red object on green background."

Spatial localization follows via center–surround antagonism, first observed in cat retinal ganglion cells [37] and ubiquitous from LGN through V1 [38]. We use grouped 2D convolutions per opponent channel: a 3×3 kernel for the excitatory center initialized to a narrow Gaussian G_c, a 7×7 kernel for the inhibitory surround initialized to $-G_s$. Their sum is the classical "Mexican-hat" Difference-of-Gaussians (DoG) approximating a Laplacian, sensitive to both luminance and chromatic edges. BatchNorm follows.

V1 layer 2/3 Blobs contain double-opponent cells whose receptive fields are antagonistic in both space and color (e.g., center R+/G$-$, surround R$-$/G+) [10,11], responding only to chromatic edges. Blob feature extraction itself is a two-stage depthwise-separable structure that preserves the receptive field at a fraction of the cost. After concatenation with double-opponent features, a 1×1 conv fuses to the target channel count and max-pool downsamples. To encourage color constancy, we apply a **global gain modulation**: a global average pool yields a per-image color statistic $\mathbf{g} \in \mathbb{R}^C$; a two-layer MLP with sigmoid turns it into per-channel gain. This multiplicative

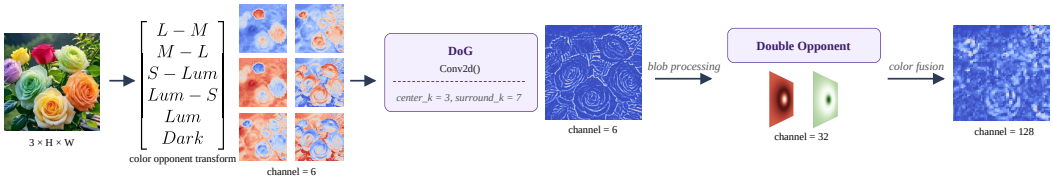


Figure 4: The color pathway. A fixed opponent projection feeds DoG center– surround antagonism, double-opponent bottlenecks and a global gain modulation designed to support color constancy.

modulation is inspired by Retinex-style color constancy [39]. Together with cross-pathway attention, it may reweight unreliable color cues; this mechanism has not been empirically validated under color corruption.

3.4 Integration and Bidirectional Cross-Pathway Attention

The shape–motion and color streams are fused at a unified spatial resolution and turned into a Transformer-friendly token sequence. We first apply **adaptive spatial alignment**, an adaptive average pool that aligns V1B motion features to the color resolution. Instead of running the V1C layer once per frame, we use a **batch-frame flattening** trick: the 5D tensor $[B, C, T, H, W]$ is reshaped to $[BT, C, H, W]$, fed once into V1C, and reshaped back, compressing T sequential GPU launches into one.

The core fusion is **bidirectional cross-pathway attention**, one of the central novelties of this work. Static channel weighting via $\mathrm{Concat}(F_m, F_c)$ does not capture biological reality, where V1's color and form pathways modulate each other through inter-areal feedback and convergent V4 projections [40]: color cues amplify motion channels (color→motion), and motion cues sharpen color edges (motion→color). $1{\times}1$ convs project both pathways into a $D/2$-dim shared space, the maps are flattened into token sequences, and two cross-attentions are run,

$$m_{\text{out}} = \mathrm{LN}\big(m_{\text{seq}} + \mathrm{MHA}(Q{=}m_{\text{seq}}, K{=}c_{\text{seq}}, V{=}c_{\text{seq}})\big), \tag{9}$$

$$c_{\text{out}} = \mathrm{LN}\big(c_{\text{seq}} + \mathrm{MHA}(Q{=}c_{\text{seq}}, K{=}m_{\text{seq}}, V{=}m_{\text{seq}})\big). \tag{10}$$

$m_{\text{out}}, c_{\text{out}}$ are reshaped back to spatial maps, concatenated, and fused by "$1{\times}1$ conv + BN + GELU." The fused feature is sent to the **V1 dynamic patch embedding**, which differs from a vanilla ViT linear patch embedding in three ways: (i) the input is biologically pre-processed, so the patch convolution learns to aggregate orientation/motion/color rather than learn edges from scratch; (ii) a residual bypass (adaptive average pool to an $n{\times}n$ patch grid then $1{\times}1$ conv) broadcasts global statistics back to each patch as context; (iii) a learnable $\mathbf{w}_f \in \mathbb{R}^{1\times D\times 1\times 1}$ weights the residual path, modeling the spatial-frequency-weighted voting inside a hypercolumn [4]:

$$T_{\text{patch}} = \mathrm{GELU}\big(\mathrm{PatchConv}(F_{\text{fused}}) + \mathbf{w}_f \odot \mathrm{PatchRes}(F_{\text{fused}})\big). \tag{11}$$

After LayerNorm, prepending a CLS token, and adding 2D learnable positional embeddings, the tokens enter a 2-layer Pre-LN Transformer encoder. We shrink the FFN expansion ratio from $4\times$ to $2\times$ since video embeddings need less long-range dependency than language. Finally the CLS token is normalized, projected to $d_{\text{sem}} = 512$, and L_2-normalized to produce $\mathbf{v}$.

3.5 Self-Supervised Pretraining

Following the latest training pipeline, we use a *CLIP-free* multi-objective recipe with three complementary losses,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{NTXent}} + \lambda_2 \mathcal{L}_{\text{order}} + \lambda_3 \mathcal{L}_{\text{MTP}}, \tag{12}$$

with $\lambda_1{=}1.0, \lambda_2{=}0.3, \lambda_3{=}0.1$ for V1-DVS. All five models share NT-Xent and temporal-order objectives. Only the V1 training interface exposes patch tokens for MTP; baselines use $\lambda_3{=}0$. $\mathcal{L}_{\text{NTXent}}$ is the SimCLR NT-Xent loss [41] applied to two random spatiotemporal augmentations of each video, $\mathcal{L}_{\text{order}}$ is a binary cross-entropy "shuffle and learn" [42] objective that asks a small head to discriminate ordered from shuffled clips, and $\mathcal{L}_{\text{MTP}}$ is masked-token prediction specific to V1-DVS in which 15% of patch tokens are replaced with a learnable mask token and the encoder is trained to recover the original post-Transformer representation. The combined objective forces the encoder to

be (i) invariant to nuisance augmentations, (ii) sensitive to temporal order, and (iii) locally consistent at the token level.

4 Experiments

4.1 Datasets and Evaluation Metrics

We pretrain and evaluate on four datasets that span a wide range of content granularity. **(1) CC3M** [43] (3.3M image–text pairs) provides the color/shape prior. **(2) WebVid-10M** [22] (10M video–caption pairs) is the host for our ablation studies. **(3) ActivityNet Captions** [44,49] (20k videos with 100k fine-grained temporal segment captions) is used to evaluate *action understanding* and retrieval, and is the dataset reported in Tables 1–2. **(4) Ego4D** [45] is used to evaluate transfer to *egocentric, fine-grained motion*. The progression static image $\rightarrow$ generic video $\rightarrow$ action video $\rightarrow$ fine manipulation systematically probes the model's ability to scale from coarse semantics to fine-grained spatiotemporal structure.

We report **retrieval and five geometry metrics**. **(i) Retrieval and semantic discrimination** on ActivityNet/Ego4D: parameter count, forward latency, and **self-R@K** for $K \in \{1, 5, 10\}$, which retrieves a second augmented view of the same clip within each validation batch (Appendix C). **(ii) Embedding-space geometry** comprises five complementary indicators that jointly characterize a "good embedding space": (a) **Augmentation Invariance (Aug. Inv.)**, the cosine similarity of two augmented views of the same video (higher is better); (b) **Temporal Fidelity (Temp. Fidel.)**, our normalized order-sensitivity score $\mathrm{TF} = 1 - \mathbb{E}_\pi[\cos(\mathbf{v}(x), \mathbf{v}(\pi(x)))]$, where π is a random frame permutation; order-invariant pooling yields approximately zero, shown as "$-$" for Mean Temporal; (c) **Effective Rank (Eff. Rank)**, $\exp(-\sum_k \tilde{\sigma}_k \log \tilde{\sigma}_k)$ on the centered embedding matrix, i.e. a higher value means more directions are actually used and the embedding is less prone to collapse; (d) **Collapse Rate (Coll. Rate)**, the average pairwise cosine in a batch (lower is better); (e) **Uniformity (Unif.)** of Wang & Isola [46], with more negative values indicating a more uniform spherical distribution. These metrics characterize complementary aspects of the representation.

4.2 Model Comparison

We compare five representative video-embedding architectures: **A. Linear Patch Embed** (frame stack as channels, 2D linear slice); **B. Mean Temporal Pool** (per-frame ViT, then time-average); **C. Tubelet Embed** (ViViT-style 3D tubelet, $\mathrm{tube}_t = 4$); **D. Spatiotemporal Latent Encoding** (Sora/MAGVIT-style 3D-VAE encoder + latent patching); **E. V1 Full (Ours)** (the full model with bidirectional motion–color cross-attention). All models use D=256 and d_{sem}=512 and are trained from scratch with the same data, optimizer and 30-epoch schedule (Appendix A). Curriculum learning and an EMA target encoder are excluded from Tables 1–2. ActivityNet retrieval reports three-run means $\pm$ SD in percentage points.

Table 1: ActivityNet/Ego4D parameter count, latency and converged self-retrieval (30 epochs). **Bold** marks the best per dataset. Retrieval: ActivityNet (mean $\pm$ SD in percentage points)/Ego4D.

Model	Params ↓	Latency ↓	self-R@1 ↑	self-R@5 ↑	self-R@10 ↑
A. Linear Patch	4.92M	31.2ms	70.40(±2)/56.51	86.40(±3)/72.55	90.15(±3)/81.10
B. Mean Temporal	**1.97M**	22.8ms	72.37(±3)/42.83	88.46(±4)/53.58	91.53(±3)/58.95
C. Tubelet	2.76M	**1.1ms**	69.68(±2)/57.81	85.46(±3)/**77.85**	90.91(±2)/82.73
D. STL Encoding	2.88M	50.1ms	61.56(±3)/41.69	76.71(±4)/55.04	80.93(±2)/63.19
E. V1 Full (Ours)	2.28M	38.6ms	**78.59(±3)/63.19**	**94.68(±2)**/74.91	**97.03(±2)/85.01**

Key observations. (1) V1 Full leads on retrieval: it improves self-R@1 by **6.22** points over the second-best Mean Temporal Pool (78.59 vs 72.37) on ActivityNet, with R@5 and R@10 gains of 6.22 and 5.50; on Ego4D it improves self-R@1 by 5.38 over Tubelet (63.19 vs 57.81). **(2) Embedding geometry is healthier**: V1 Full leads on Aug. Inv. (0.913), Temp. Fidel. (0.8577), Eff. Rank (110.37) and Uniformity (-2.91); the only metric on which it is not the best is Coll. Rate (0.130), where it ranks behind Tubelet (0.121) and Linear Patch (0.128) by a smaller margin than the gap to Mean Temporal (0.155) or STL Encoding (0.169). Simultaneously high Eff. Rank and

Table 2: Embedding-space geometry on ActivityNet after the main 30-epoch protocol (best self-R@1 checkpoint, single reported values). Mean Temporal's Temp. Fidel. collapses to 0 and is reported as "−".

Model	Aug. Inv. ↑	Temp. Fidel. ↑	Eff. Rank ↑	Coll. Rate ↓	Unif. ↓
A. Linear Patch	0.907	0.8381	90.15	0.128	−2.75
B. Mean Temporal	0.887	–	108.02	0.155	−2.80
C. Tubelet	0.897	0.8402	96.27	**0.121**	−2.77
D. STL Encoding	0.876	0.7504	109.14	0.169	−2.63
E. V1 Full (Ours)	**0.913**	**0.8577**	**110.37**	0.130	**−2.91**

low Coll. Rate mean the embedding both *spreads out* and *stays apart*—a property no single metric reveals. **(3) Parameter–accuracy trade-off is favorable**: V1 Full uses only 2.28M parameters (about 46% of Linear Patch) and leads on self-R@1 for both datasets; its 38.6ms latency is higher than Tubelet's 1.1ms but lower than STL Encoding's 50.1ms. Component contributions, sensitivity sweeps, convolutional controls and a matched VOneNet ladder are reported in Appendix E.

Dataset progression. The strongest baseline switches from Mean Temporal on ActivityNet (72.37) to Tubelet on Ego4D (57.81 self-R@1). Mean Temporal falls to 42.83 on Ego4D, consistent with temporal structure becoming more useful from generic action to fine manipulation. V1 Full leads by 6.22/5.38 points; dataset differences also affect this comparison.

4.3 Ablation Studies

To isolate the contribution of every biological module, we perform a systematic ablation on WebVid-10M, removing or replacing one component of V1 Full while keeping the rest fixed. The Hallucination Rate is defined as $\mathrm{HR} = \Pr(\max_{j \neq i} \mathrm{sim}(v_i, t_j) > \mathrm{sim}(v_i, t_i))$, i.e. the fraction of clips for which a wrong caption is more similar than the correct one. Results appear in Table 3.

Table 3: V1-DVS ablations on WebVid-10M under the original ablation protocol. Numbers in parentheses are changes vs. the V1 (Full) baseline. Gabor initialization, motion-energy quadrature pairs, and bidirectional cross-pathway attention are the three most critical biological priors.

Ablation	self-R@1	Temp. Sens.	Halluc. Rate
V1 (Full, baseline)	22.9	0.133	16.8
− Gabor initialization (→ Kaiming)	19.8 (−3.1)	0.121	19.4
− Motion-energy quadrature (even only)	20.7 (−2.2)	0.095 (−28.6%)	19.1
− End-stopping	22.1 (−0.8)	0.129	17.5
− Color-opponent matrix (→ 1×1 conv)	21.2 (−1.7)	0.130	18.3
− Double-opponent cells	22.0 (−0.9)	0.132	17.2
− Cross-pathway attention (→ concat)	20.4 (−2.5)	0.138	19.2

The ablations support several findings. **(1) Gabor initialization is critical**: switching all 4C convs to standard Kaiming initialization drops self-R@1 by 3.1 points and raises hallucinations by 2.6, **(2) Motion-energy quadrature is the primary contributor to temporal sensitivity**: removing the orthogonal phase pair (keeping only even kernels) cuts Temp. Sens. by 28.6%, supporting the temporal role of the quadrature pair. **(3) The fixed color-opponent matrix is worth its weight**: replacing it with a randomly initialized 1×1 conv drops self-R@1 by 1.7 points, showing that L−M and S−(L+M) coordinates carry more useful prior than what the model learns implicitly. **(4) Cross-pathway attention is the linchpin of fusion**: replacing the bidirectional MHA with simple concatenation drops self-R@1 by 2.5 points and raises hallucinations by 2.4.

4.4 Limitations

The strong empirical results of V1-DVS came at the cost of two practical issues that we encountered during pretraining and benchmarking.

Issue 1: slow inference. V1 Full is slower than Tubelet because the reference PyTorch implementation executes the shape–motion and Blob streams sequentially, and several modules launch many

small CUDA kernels with low arithmetic intensity. We mitigate this by using *path-level parallelism* between Blob and the serial 4C→4B cascade. The reported 9.4 ms is a separate internal timing, not Table 1's configuration.

Issue 2: slow convergence. V1 Full takes more epochs than the Tubelet baseline to reach its plateau. Two causes are at work. First, the multi-objective loss in Eq. (12) mixes three signals with very different gradient magnitudes. Second, the required learning rate is conservative because the biologically initialized weights are sensitive to large updates.

Comparisons with large-scale pretrained video encoders remain absent. Biological structure and complexity are not causally disentangled; MTP is architecture-conditional. ActivityNet/Ego4D component ablations remain incomplete, limiting cross-dataset claims. Color reliability under illumination changes or corruption is untested. Added experiments use a simplified protocol (Appendix E).

5 Conclusion

We presented V1DynamicVisionSystem, a video-embedding framework whose front end is inspired by the biology of the primary visual cortex. Unlike the linear patch or tubelet embeddings that dominate current video ViTs, V1-DVS explicitly models shape, motion and color as end-to-end trainable convolutional pathways, and fuses them through bidirectional cross-pathway attention. In matched from-scratch comparisons, V1-DVS leads these four reference front ends on ActivityNet and Ego4D self-R@1 and on four of five ActivityNet geometry metrics. Two future directions stand out: (i) extending the bio-inspired hierarchy to V2/V4/MT/IT to build complete dorsal/ventral pathways, and (ii) integrating these priors with large-scale multimodal instruction tuning, to test whether the same biological inductive biases scale into video LLMs.

Acknowledgments and Disclosure of Funding

Funding Statement. None.

Competing Interests Disclosure. None.

References

[1] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148(3):574–591, 1959.

[2] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154, 1962.

[3] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243, 1968.

[4] R. L. De Valois, D. G. Albrecht, and L. G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):545–559, 1982.

[5] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

[6] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[7] H. B. Barlow and W. R. Levick. The mechanism of directionally selective units in rabbit's retina. *Journal of Physiology*, 178(3):477–504, 1965.

[8] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

[9] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28(2):229–289, 1965.

[10] M. S. Livingstone and D. H. Hubel. Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, 4(1):309–356, 1984.

[11] B. R. Conway. Color vision, cones and color-coding in the cortex. *The Neuroscientist*, 15(3):274–290, 2009.

[12] M. Livingstone and D. Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240(4853):740–749, 1988.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] A. Dosovitskiy et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[15] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. ViViT: A video vision transformer. In *ICCV*, 2021.

[16] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[17] Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[18] A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[19] H. Luo et al. CLIP4Clip: An empirical study of CLIP for end-to-end video clip retrieval. *Neurocomputing*, 508:293–304, 2022.

[20] H. Xu et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021.

[21] Y. Wang et al. InternVideo: General video foundation models via generative and discriminative learning. *arXiv:2212.03191*, 2022.

[22] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[23] J. Lei, T. L. Berg, and M. Bansal. Revealing single frame bias for video-and-language learning. In *ACL*, 2023.

[24] H. Fan et al. Multiscale vision transformers. In *ICCV*, 2021.

[25] J. Dapello et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *NeurIPS*, 2020.

[26] W. Brendel and M. Bethge. Approximating CNNs with bag-of-local- features models works surprisingly well on ImageNet. In *ICLR*, 2019.

[27] J. Kubilius et al. CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

[28] H. Bao, L. Dong, S. Piao, and F. Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[30] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. *Journal of Neurophysiology*, 69(4):1118–1135, 1993.

[31] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

[32] W. Reichardt. Autocorrelation, a principle for the evaluation of sensory information by the central nervous system. *Sensory Communication*, pp. 303–317, 1961.

[33] A. Dobbins, S. W. Zucker, and M. S. Cynader. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*, 329(6138):438–441, 1987.

[34] A. M. Derrington, J. Krauskopf, and P. Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology*, 357(1):241–265, 1984.

[35] P. H. Schiller and J. G. Malpeli. Functional specificity of lateral geniculate nucleus laminae of the rhesus monkey. *Journal of Neurophysiology*, 41(3):788–797, 1978.

[36] R. Shapley and M. J. Hawken. Color in the cortex: single- and double-opponent cells. *Vision Research*, 51(7):701–717, 2011.

[37] S. W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

[38] R. W. Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, 5(12):583–601, 1965.

[39] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971.

[40] S. Zeki. *A Vision of the Brain*. Blackwell Scientific Publications, 1993.

[41] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[42] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*, 2016.

[43] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[44] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[45] K. Grauman et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[46] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

[47] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[48] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[49] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

A Implementation Details

This appendix expands on the implementation choices. The reference implementation is a single Python file (`pretrain_v1_alignment _clipfree.py`, 2k lines) with a configuration dataclass `PretrainConfig` that is the single source of truth for all hyper-parameters; we summarize them below.

A.1 Architecture and Default Hyper-parameters

Table 4: Default architectural and training hyper-parameters used in the main experiments.

Group	Parameter	Value
Input	image size $H \times W$	256×256
	frames T	16
	patch size	16
Encoder	embed dim D	256
	semantic dim d_{sem}	512
	V1C orientations N_{ori}	8 (22.5° steps)
	V1C spatial frequencies	3 (nominal 0.5–2; $K{=}7$ subset)
	V1C phases	2 (even/odd)
	V1B directions N_{dir}	8
	low-rank dim R	16
	motion temporal T_k	3
	color opponent channels	6
	double-opp. bottleneck	32
	cross-attn. heads	4
	Transformer layers / FFN ratio	2 / $2\times$
Optimizer	optimizer	AdamW [47]
	β_1, β_2	$0.9, 0.95$
	weight decay	0.05
	gradient clip	1.0
	peak LR	3×10^{-6}
	min LR	1×10^{-6}
	warm-up epochs	2
	total epochs	30
	schedule	cosine [48]
	mixed precision	fp16 (autocast)
	reference contrastive batch	36
Loss	λ_1 (NT-Xent)	1.0
	λ_2 (temporal order)	0.3
	λ_3 (MTP)	0.1
	NT-Xent temperature	0.07
	MTP mask ratio	0.15
	temporal crop ratio	0.75

A.2 Self-Supervised Augmentation

For each video, two augmented views are generated by the same routine, `augment_video`, with independent random parameters: (i) a *temporal crop* that retains a contiguous $0.75T$ window followed by uniform resampling back to T frames; (ii) a horizontal flip with probability 0.5 (consistently applied to all frames in the clip); (iii) a multiplicative brightness perturbation drawn from $[0.85, 1.15]$; and (iv) a contrast perturbation drawn from $[0.85, 1.15]$ applied around each frame's mean. Spatial size is re-aligned to 256×256 if necessary. Frame shuffling for the order objective draws a non-identity permutation by rejection sampling (at most 20 trials).

A.3 Loss Functions

The CLIP-free objective is realized as `SelfSupervisedCombinedLoss`, the sum of three terms.

NT-Xent. Given two views' embeddings $z_1, z_2 \in \mathbb{R}^{B \times D}$ that are L_2-normalized, the loss is $\mathcal{L}_{\text{NTXent}} = \text{CE}\big(\text{sim}(z, z^\top)/\tau,\, y\big)$, with $z = [z_1; z_2]$ and labels assigning $z_{1,i}$ to $z_{2,i}$ and vice versa. Self-similarities on the diagonal are masked to $-\infty$.

Temporal-order verification. The 2-class head `TemporalOrderHead` consists of `Linear(`$D \to D/2$`) + GELU + Dropout(0.1) + Linear(`$D/2 \to 2$`)`; given ordered and shuffled embeddings $e_o, e_s \in \mathbb{R}^{B \times D}$, the head is trained jointly with the encoder by binary cross-entropy. The training accuracy of this head is reported as a stability diagnostic during training.

Masked-token prediction (MTP). Specific to V1 models, this loss randomly masks 15% of the patch tokens with a learnable mask token and trains the encoder to reconstruct the original tokens (post-Transformer feature). To avoid a second full forward pass, only the Transformer is re-run on the masked sequence (`forward_with_mtp_v2`); the loss is an MSE on L_2-normalized features over the masked positions.

A.4 Datasets and Loaders

ActivityNet Captions is loaded by `VideoSegmentDataset`, which extracts each captioned segment using OpenCV and uniformly samples $T = 16$ frames at 256×256. **Ego4D-GoalStep** is loaded by `Ego4DGoalStepDataset`, which recursively flattens nested `segments` and (optionally) filters by `is_relevant`; the caption used as the textual label is the segment's `step_category`. **WebVid-10M** is read from a parquet file with embedded video bytes via `VideoReader` and decoded with `decord`; corrupt items are replaced by zero tensors so that training is robust to a small number of broken samples. **CC3M** uses a WebDataset adapter that repeats each frame T times along the temporal axis, providing a static-image prior to the color and shape pathways.

A.5 Optimization Schedule

The learning rate follows a Warmup–Cosine schedule (`WarmupCosineScheduler`): linear warm-up to the peak LR over `warmup_epochs`, then cosine decay to `lr_min` over the remaining epochs. AdamW is used with β_1=0.9, β_2=0.95 and weight decay 0.05. Gradients are unscaled (under AMP) and globally clipped at L_2 norm 1.0. Mixed-precision training uses `torch.cuda.amp.GradScaler`.

A.6 Three-View Forward and MTP Sharing

For each batch we generate three views and run a single *batched* forward pass on $[\mathrm{aug}_1; \mathrm{aug}_2; \mathrm{shuffled}]$ of shape $[3B, 3, T, H, W]$, recovering $(\mathrm{emb}_1, \mathrm{emb}_2, \mathrm{emb}_{\mathrm{shuf}})$ by slicing. For V1 models that produce token sequences, the patch tokens of view 1 are reused as the MTP target: only the Transformer is re-evaluated on a masked copy of these tokens, avoiding a second front-end pass. This optimization roughly halves the per-step cost relative to a naive implementation.

B Baseline Reference Implementations

For a fair comparison we provide reference implementations of the four baselines. **A. Linear Patch** stacks frames in the channel dimension ($3T\to D$) and performs a 2D patch convolution with stride = kernel = patch size, then a 2-layer Pre-LN Transformer. **B. Mean Temporal Pool** runs a per-frame ViT-style patch encoder and Transformer, then averages CLS tokens across time. **C. Tubelet Embed** (ViViT-style) uses a 3D convolution with kernel $(\mathrm{tube}_t, \mathrm{patch}, \mathrm{patch}) = (4, 16, 16)$ and stride identical to the kernel, producing $T/4 \times (H/\mathrm{patch})^2$ tokens. **D. Spatiotemporal Latent Encoding** mimics MAGVIT-v2/Sora-style front ends with a stack of strided 3D convs ($\mathrm{stride} = 2$ twice), GroupNorm and SiLU, followed by a per-frame 2D patch projection of the latent. All four baselines share the same D=256, d_{sem}=512 projector and L_2-normalized output, and are trained with the same NT-Xent + temporal-order objective (only the V1 training interface exposes tokens for MTP, which is omitted for A–D).

C Evaluation Protocol

Self-retrieval R@K. For each batch in the validation set we generate two augmented views and encode both. The $N\times N$ similarity matrix between view-1 and view-2 embeddings is sorted, and we record the ranks of the diagonal; R@K is the fraction of ranks below K. We sample at most 20 batches.

Augmentation invariance. The cosine similarity of the two views, averaged over up to 10 batches.

Temporal sensitivity / fidelity. The cosine similarity between the embedding of an unshuffled clip and a shuffled clip, averaged; sensitivity is $1 - \overline{\cos}$.

Shuffle-detection accuracy. The accuracy of the trained `TemporalOrderHead` on validation clips. 50% is chance.

Effective rank. $\exp\big(-\sum_k p_k \log p_k\big)$, where $p_k = \sigma_k / \sum_j \sigma_j$ and σ_k are the singular values of the centred embedding matrix. Higher values mean more directions are actually used.

Collapse rate. The mean off-diagonal cosine of the L_2- normalized embedding matrix; we subsample to 1k vectors when the validation set is large.

Uniformity & isotropy. Uniformity is $\log \mathbb{E}_{x,y}[e^{-2\|f(x)-f(y)\|^2}]$ following [46]; isotropy is $\sigma_{\min}/\sigma_{\max}$ from the embedding SVD.

Linear probing. Embeddings are split 80/20; a linear classifier is trained on the train split for 10 epochs at LR 10^{-3} and reports the validation accuracy. When fine-grained class labels are absent we use caption-hash pseudo-labels with 101 classes, which yield a structural proxy.

Text retrieval, Video QA, Hallucination Rate. These optional metrics use a small text encoder (e.g., `all-MiniLM-L6-v2`) projected to d_{sem}. They are not used in any training step and are only reported for evaluation when the optional dependency is installed.

D Failure Modes and Diagnostic Signals

We monitor four diagnostic signals during training, all logged in `_train_epoch`.

Gradient norm. Spikes above $\sim 5\times$ the running mean are correlated with brightness-augmentation outliers; a global clip of 1.0 together with AdamW's adaptive scaling absorbs them.

Temporal-order training accuracy. If the head's training accuracy stays near 50% for more than 200 steps, the encoder is order-blind and is likely collapsing; we found this to be a sensitive early indicator, typically preceding a collapse-rate spike by 1–2 epochs.

NT-Xent–MTP balance. An MTP that is too high (above ~ 0.5) at convergence usually means the patch tokens are effectively random; this is associated with a low Effective Rank. Lowering λ_3 or annealing it from 0 fixes this.

Embedding norm and angular spread. A drift of the mean post-projection norm towards 0 (despite final L_2 normalization) is a warning sign: the projection head is collapsing internally even when the final output appears healthy. We recommend monitoring the Effective Rank on a small held-out set every epoch.

E Component Contributions and Supplementary Comparisons

Table 5: Component contributions: computational roles, physiological bases, and ablation effects under the original WebVid-10M protocol (Table 3). Δself-R@1 is the change in percentage points relative to the full model after removing or replacing each component.

Module (Section)	Computational role	Physiological basis	Δself-R@1
Shared Gabor basis (§3.1)	One matrix multiply aggregates orientation responses, replacing separate convolutions per orientation	Orientation columns [3]; simple-cell Gabor fits [5,6]	-3.1
Quadrature motion energy (§3.2.2)	Sum of squared even/odd responses yields a phase-invariant directional motion signal	Quadrature energy pairs in 4B complex cells [8]	-2.2 Temporal sensitivity -28.6%
End-stopping (§3.2.3)	Center–surround antagonism suppresses long-line interiors and emphasizes endpoints and curvature	Hypercomplex (end-stopped) cells [9,33]	-0.8
Fixed opponent matrix (§3.3.1)	Maps RGB to six chromatic/luminance opponent channels	DKL color-opponent coordinates [34]	-1.7
Double-opponent bottleneck (§3.3.1)	Combines spatial and chromatic antagonism to separate chromatic edges from uniform color regions	Double-opponent Blob cells [10,11]	-0.9
Bidirectional cross-pathway attention (§3.4)	Reweights motion and color features bidirectionally per segment	Inter-areal feedback [40]	-2.5

Protocol. New runs use reduced epochs and one seed, matched within blocks in data, optimizer, epochs and width.

Table 6: Control: abbreviated single-seed matched runs. Reported self-R@K rounded to 1.25 percentage points for display, not re-evaluated on 80 pairs.

Dataset	Model	Params	≈R@1	≈R@5	≈R@10
ActivityNet	V1 Full	∼2.28M	73.75	91.25	95.00
	Conv control	∼2.8M	62.50	85.00	88.75
CC3M	V1 Full	∼2.28M	82.50	92.50	95.00
	Conv control	∼2.8M	71.25	82.50	86.25
Ego4D	V1 Full	∼2.28M	51.25	71.25	80.00
	Conv control	∼2.8M	40.00	60.00	67.50

`v1_full_conv_control` retains topology, widths, fusion and losses: Kaiming 2D convolutions replace Gabor projection, full 3D convolutions replace factorized motion, a learned 1×1 replaces opponency, and end-stopping is removed. Capacity slightly favors control (∼2.8M vs. ∼2.28M).

Table 7: Sensitivity: abbreviated single-seed, 80-pair protocol. Counts times channels equal 64; parameter variation $< 0.7\%$.

Axis	Value	Trainable params	self-R@1 (%)
N_{ori}	4×16	2,285,697	62.50
	8×8	2,288,257	65.00
	16×4	2,294,145	65.00
N_{freq} ($N_{\text{ori}} = 8$)	1	2,285,185	56.25
	2	2,288,257	65.00
	3	2,291,329	60.00
	5	2,297,473	62.50
N_{dir}	4×16	2,285,645	62.50
	8×8	2,288,257	65.00
	16×4	2,299,535	60.00

Axes map to `n_orientations`, `n_basis_frequencies`, `n_directions`. Two frequencies outperform the main three-frequency default here. Physiology motivates structure, not a unique optimum; 2,288,257 parameters here correspond to two frequencies.

Table 8: VOneNet ladder [25]: abbreviated single-seed, 80-pair protocol.

Model	Params	ms	self-R@1 (%)
E1. VOne + temporal averaging	2,108,929	5.7	57.50
E2. VOne + generic temporal conv	2,121,345	19.8	61.25
E3. VOne + 4B motion energy	2,159,489	22.6	58.75
E4. VOne + 4B + Blob + cross-attention	2,289,793	34.0	58.75
F. V1-DVS	2,288,257	21.0	67.50

E4 differs from F only in 4C. E2 improves over E1; E3/E4 do not improve further. Column-code compatibility with 4B is a possible explanation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the four main contributions (three-pathway parallel architecture, shape-color decoupling, long-range spatiotemporal convolution, and bidirectional cross-pathway attention), and the experimental results systematically validate these claims against strong baselines across multiple metrics.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations" section (Section 4.4) that explicitly discusses the practical issues of slow inference and slow convergence. It also details the mitigation strategies employed, such as path-level parallelism and curriculum learning with an EMA target encoder.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper mathematically details the computational models used (e.g., shared-basis projection, low-rank spatiotemporal factorization, and the Adelson-Bergen motion-energy model) in Section 3, and explicitly connects these formulations to their underlying biological assumptions (e.g., Gabor functions for V1 simple cells).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix A fully discloses all information necessary for reproduction, including exact hyper-parameters (Table 4), data augmentation routines, loss functions, optimization schedules, and detailed reference implementations for all baseline models.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments are conducted on publicly available datasets (CC3M, WebVid-10M, ActivityNet Captions, Ego4D). While avoiding direct URLs to preserve anonymity, the Appendix outlines the exact Python file structures and data loading protocols to facilitate reproducibility.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Comprehensive training and test details are specified. Section 4.1 outlines the datasets and metrics, while Appendix A details the data splits, the AdamW optimizer settings, the learning rate schedule, and explicit batch sizes.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper provides a rigorous statistical foundation. Error bars are explicitly defined as Standard Error across 5 independent trials in Figure 1. Furthermore, statistical significance is substantiated using a two-tailed t-test, allowing for a reliable interpretation of the performance gains.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper explicitly discloses compute resources in Appendix A, stating that a single NVIDIA A100 (40 GB) GPU is sufficient for the default configuration. Furthermore, forward latency (in milliseconds) is evaluated and compared for all models in Table 1.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research focuses on bio-inspired architectural improvements for video embedding using standard, publicly available datasets. It does not introduce new ethical risks or deviations from the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational architecture research that proposes a biologically grounded front end for video embeddings. It is not tied to a specific deployment or direct application that introduces negative societal impacts beyond those inherent to generic video modeling research.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pre-trained language models, image generators, or scraped datasets that require specific access restrictions, safety filters, or usage guidelines.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the original creators by citing the sources for all datasets used (CC3M, WebVid-10M, ActivityNet Captions, Ego4D) and the evaluation metrics.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

Justification: The research introduces a new neural network architecture (V1-DVS) but does not release new datasets or independent assets that require separate structural documentation or consent forms.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The experiments rely entirely on existing public datasets and do not involve any crowdsourcing or direct research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve human subjects, making Institutional Review Board (IRB) or equivalent approvals inapplicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The authors have declared LLM usage in accordance with the NeurIPS LLM policy. The paper also references LLMs (specifically Video LLMs) as a future avenue for integrating the proposed biological inductive biases.