International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 263-277
https://doi.org/10.70917/ijcisim-2026-5179
Article

TE-CUME: Uncertainty-Aware Tri-Modal Video Retrieval Representation Learning

Sanjib Kumar Swain[1], Santosh Kumar Swain[2]

[1]School of Computer Engineering, KIIT University, Bhubaneswar, India 2081027@kiit.ac.in
[2]School of Computer Engineering, KIIT University, Bhubaneswar, India sswainfcs@kiit.ac.in

Abstract: Automatic Content Recognition (ACR) systems deployed in smart TVs rely on multimodal encoders that produce deterministic fingerprint embeddings, offering no mechanism to express how trustworthy a particular identification is under varying broadcast degradation. This paper introduces TE-CUME (Tri-Encoder Cross-Modal Uncertainty Fusion Model), a probabilistic tri-modal encoder that equips each modality stream — visual (ViViT-S), audio (BEATs), and on-screen text (TrOCR) — with Monte Carlo Dropout and heteroscedastic output heads, producing per-modality Gaussian posteriors over the content embedding space. A minimum-variance Product-of-Experts fusion stage combines these posteriors through precision weighting, provably minimising the fused estimate's variance. We propose UAT-InfoNCE, a precision-weighted contrastive objective that adaptively modulates gradient contributions by the model's own confidence, together with pairwise cross-modal disagreement terms (,,) that serve as zero-shot diagnostics for broadcast failure modes such as A/V desynchronisation, dubbed content, and advertisement splices. Evaluation on ActivityNet Captions focuses primarily on video-text retrieval as a direct encoder-quality benchmark across the full 4,917-video validation set (R@, mAP, ECE), complemented by a conceptual architectural specification showing how TE-CUME serves as a modular Video Feature Extractor (VFE) replacement for dense video captioning pipelines (PDVC, Vid2Seq) while propagating calibrated per-event confidence scores. Experimental results demonstrate that predicted epistemic uncertainty correlates positively with degradation severity (Spearman,), and that TE-CUME achieves competitive retrieval performance (R@1 = 43.1 on ActivityNet) while uniquely providing per-query calibrated confidence that enables downstream systems to act conservatively on unreliable representations.
Keywords: Video Representation Learning, Uncertainty Quantification, Multimodal Fusion, Dense Video Captioning, Monte Carlo Dropout, Calibration, Video-Text Retrieval

1. Introduction

Automatic Content Recognition (ACR) has emerged as a critical technology in the smart TV ecosystem, enabling platforms to identify what content is being displayed at any given moment by periodically capturing frames and audio, building a fingerprint, and matching it against a content library [1], [2]. Since its inception in 2011, with roots in Shazam-like audio identification, ACR has been adopted by all major smart TV manufacturers — Samsung, LG, Vizio, and Roku — and is now deployed in nearly three-quarters of connected television households [1]. At a high level, as illustrated in Fig., the ACR pipeline consists of a client-side encoder that extracts multimodal fingerprints from the displayed content, and a server-side matching engine that identifies the content against a reference database.

The encoder component is the computational core of this pipeline: it must produce discriminative representations from video, audio, and on-screen text streams that are robust to the diverse degradation conditions encountered in real-world broadcast environments. Modern learned encoders [3], [4] have demonstrated strong performance under controlled conditions, and a comprehensive survey of 74 dense video captioning methods [5] confirms that feature extraction has advanced significantly since the 2017 ActivityNet challenge [6]. However, a fundamental limitation persists across these approaches: existing multimodal encoders produce *deterministic* embeddings with no associated confidence estimate.

Copyright: © 2026 by the authors

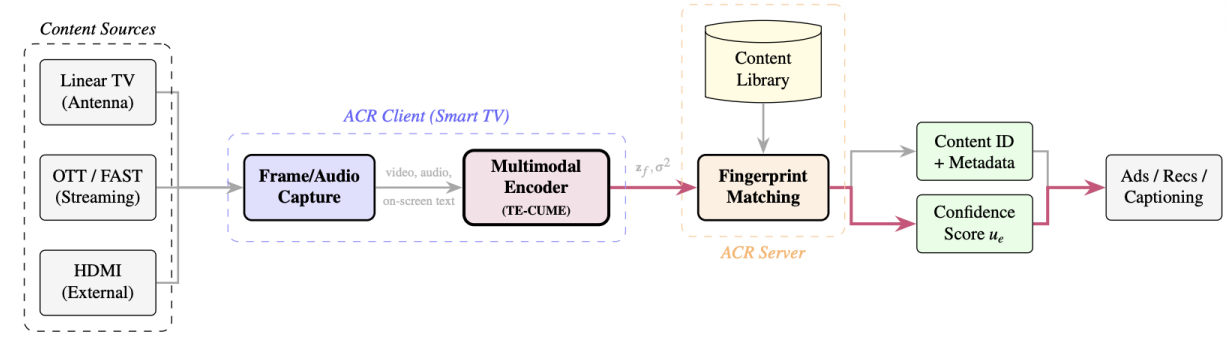


Fig. 1. Overview of the ACR pipeline. Content from diverse sources (linear TV, OTT streaming, HDMI input) is periodically captured by the ACR client embedded in the smart TV's operating system. The multimodal encoder (highlighted; replaced by TE-CUME in this work) extracts a fingerprint embedding $\mathbf{z}_f$ together with a calibrated uncertainty estimate σ^2. The ACR server matches fingerprints against a content library and returns a content identifier. Unlike conventional systems, TE-CUME additionally propagates a per-query confidence score u_e to downstream applications (advertising, recommendations, captioning), enabling them to act conservatively on unreliable identifications.

In practice, input quality varies widely across broadcast scenarios: lossy compression (H.264/HEVC at varying bitrates), additive channel noise, perspective distortion from phone-of-TV capture, station-logo overlays, and temporal dropout are commonplace [7]. Under such conditions, representation quality degrades in a query-dependent manner that the encoder cannot communicate. A downstream system consuming deterministic embeddings — whether for retrieval, recommendation, audience measurement, or dense video captioning — treats every fingerprint identically, even when the underlying signal falls close to the noise floor.

This absence of calibrated confidence is consequential across the full ACR application stack. In content identification, a silent mis-match propagates into incorrect audience measurement or financial settlement between content owners and distributors. In dense video captioning, an encoder that cannot signal ambiguity may cause the decoder to hallucinate descriptions for corrupted segments [5]. In recommendation, low-confidence identifications can trigger irrelevant suggestions that degrade user trust. These failure modes motivate the core thesis of this paper: *a calibrated, decomposable uncertainty estimate has value independent of any improvement in raw accuracy*, because it enables every downstream consumer to act differently on reliable versus unreliable representations.

We identify three specific limitations of existing multimodal encoders that prevent uncertainty-aware operation:

1. **Fixed-weight or attention-gated fusion.** Existing methods that combine modalities [2] assign fixed or attention-learned weights rather than weighting by each modality's *current* reliability under the observed degradation. When one modality is corrupted (e.g., audio dropout during an HDMI glitch), it contributes equally to the fused representation, diluting rather than preserving the information from intact streams.
2. **Deterministic representations.** Neural encoders produce point embeddings without an associated uncertainty estimate, collapsing the distinction between epistemic uncertainty (model uncertainty, reducible with more data) and aleatoric uncertainty (irreducible observation noise from the capture environment).
3. **No downstream-propagable confidence.** Neither state-of-the-art retrieval systems (CLIP4Clip [4], InternVideo2 [3]) nor dense video captioning architectures (PDVC [8], Vid2Seq [9]) report calibration metrics (ECE) or per-event confidence scores, leaving the reliability of any implicit quality signal entirely unknown to downstream consumers.

Proposed approach. To address these limitations, we present **TE-CUME** (Tri-Encoder Cross-Modal Uncertainty Fusion Model), a probabilistic multimodal encoder designed for uncertainty-aware video representation learning. TE-CUME instruments three pre-trained transformer backbones — ViViT-S [10] for spatio-temporal visual features, BEATs [11] for audio, and TrOCR [12] for on-screen text recognition — with Monte Carlo Dropout [13] and heteroscedastic output heads that produce per-modality Gaussian posteriors over the content embedding space. These posteriors are combined through a minimum-variance Product-of-Experts (PoE) fusion stage that provably allocates higher weight to more confident modalities. Training employs a novel precision-weighted contrastive objective (UAT-InfoNCE) that adaptively modulates gradient contributions by estimated fused precision. Critically, TE-CUME is designed as a *drop-in encoder*: it can replace the feature-extraction stage of any existing dense video captioning or retrieval pipeline without modifying proposal generation or decoding components.

Evaluation strategy. We evaluate TE-CUME on ActivityNet Captions [6] focusing on representation quality and calibration:

- Video-text retrieval (primary) — a direct encoder-to-encoder benchmark measuring representation quality, calibration, and noise resilience via R@, mAP, ECE,, and Spearman between predicted uncertainty and retrieval error across the full 4,917-video validation set.
- **Dense video captioning specification** (secondary/architectural) — a modular drop-in specification framing TE-CUME as a Video Feature Extractor (VFE) replacement for downstream captioning pipelines (PDVC, Vid2Seq) without modifying proposal generation or decoding blocks.

Contributions. In summary, our contributions are as follows:

1. **A probabilistic tri-modal encoder (TE-CUME)** that produces calibrated, decomposable per-modality uncertainty via Product-of-Experts precision pooling, serving as a drop-in replacement for deterministic encoders in video representation learning pipelines.
2. An uncertainty-aware contrastive training objective (UAT-InfoNCE) that weights gradient contributions by estimated fused precision, together with three pairwise cross-modal disagreement terms (,,) that serve as zero-shot diagnostics for broadcast failure modes (A/V desynchronisation, dubbed audio, advertisement splices).
3. A rigorous evaluation protocol focused on video-text retrieval and calibration under degradation, introduceing the Uncertainty-Conditional mAP metric to evaluate whether predicted confidence is actionable, complemented by an architectural integration specification for dense captioning.

2. Related Work

2.1 ACR and Video Fingerprinting

The ACR pipeline has evolved from perceptual hashing [14] — binary codes matched via Hamming distance — through CNN-based descriptors [15], [16] to transformer embeddings. The VCSL challenge [7] introduced large-scale evaluation for temporal copy detection with segment-level metrics (SegAP, Detection Precision), with subsequent work on hierarchical transformers [17] and fine-grained alignment [18] pushing accuracy on this benchmark. Related benchmarks include FIVR-200K [19], VCDB [20], and CC_WEB_VIDEO [21].

A consistent limitation across this line of work is the output contract: every system emits a distance (or similarity score) to be thresholded, with no mechanism for expressing how trustworthy that score is for a particular query. The decision boundary is global, yet the degradation regime is query-specific. TE-CUME addresses this by replacing the point embedding with a distributional output whose variance serves as an intrinsic per-query confidence signal.

2.2 Multimodal Fusion for Video Understanding

Transformer-based video encoders (TimeSformer [22], ViViT [10], VideoMAE [23]) provide strong spatio-temporal representations. BEATs [11] offers a competitive self-supervised audio encoder, and TrOCR [12] provides

robust scene-text recognition. Multimodal fusion strategies range from simple concatenation and mean-pooling to learned gating mechanisms [2].

However, existing fusion methods for video retrieval and ACR do not distinguish between a modality that is confidently informative and one that is degraded, because they operate on deterministic representations. TE-CUME makes this distinction explicit by fusing distributional estimates, so that a degraded modality automatically contributes less to the fused representation through the precision-weighting mechanism.

2.3 Uncertainty Quantification in Deep Learning

Bayesian deep learning [24] decomposes predictive uncertainty into epistemic (model) and aleatoric (data) components. MC-Dropout [13] provides a practical approximation to epistemic uncertainty through stochastic forward passes. Deep ensembles [25] offer well-calibrated estimates at the cost of multiple model copies. Temperature scaling [26] applies post-hoc calibration. Evidential deep learning [27] produces single-pass uncertainty via higher-order distributions.

These methods have been applied extensively in classification, segmentation, and active learning. Recent work has begun to explore uncertainty in retrieval: probabilistic embeddings for cross-modal matching [28], probabilistic representations for video contrastive learning [29], and benchmarks for transferable uncertainty estimates [30]. In particular, Probabilistic Cross-Modal Embedding (PCME) [28] introduced soft contrastive learning for bi-modal image-text matching by embedding inputs as Gaussian distributions to capture 1-to-many match ambiguity. However, PCME focuses exclusively on bi-modal image-text settings and assumes fixed observation noise without decomposing uncertainty into epistemic and aleatoric components or providing cross-modal disagreement diagnostics. Temperature scaling has also been studied specifically in the context of contrastive objectives [31]. TE-CUME extends probabilistic representation learning to tri-modal video-audio-text systems under broadcast degradation, combining MC-Dropout with heteroscedastic heads, precision-weighted Product-of-Experts fusion, and zero-shot cross-modal disagreement terms. We treat PCME, deep ensembles, temperature scaling, and evidential learning as natural calibration baselines for comparison (Section).

2.4 Dense Video Captioning

Dense video captioning (DVC) jointly localises temporal events and generates per-event natural-language descriptions. A recent comprehensive survey [5] catalogues 74 DVC methods (2018–2023), organising them into a canonical three-stage pipeline: (1) Video Feature Extraction (VFE), (2) Temporal Event Localisation (TEL), and (3) Dense Caption Generation (DCG). Feature backbones have progressed from C3D and I3D through CLIP to Vid2Seq's [9] language-model–based token prediction. Temporal localisation spans sliding-window (DAP, SST), boundary-aware (BSN, BMN), and DETR-style parallel decoding (PDVC [8]). Caption decoders include hierarchical LSTMs, masked transformers [32], and unified encoder–decoder architectures (VLCAP, VLTinT).

Crucially, across all 74 methods reviewed, none provides a calibrated per-event confidence estimate or reports uncertainty-aware evaluation metrics (ECE, Spearman). The survey explicitly identifies "encoder limitations making the decoder short-sighted" and the need for "diverse feature fusion" as open challenges [5]. TE-CUME directly addresses both: it replaces the VFE stage with a probabilistic encoder that communicates per-representation reliability to downstream TEL and DCG components through calibrated uncertainty.

2.5 Contrastive Learning for Retrieval

InfoNCE and its supervised variants [33] form the standard training paradigm for metric learning. These objectives weight all batch elements equally regardless of the model's current confidence in them. A difficulty-aware variant, curriculum-InfoNCE, schedules samples by a fixed difficulty proxy, but does not use the model's own uncertainty as the weighting signal. Concurrent work on probabilistic embeddings [28], [29] learns distributional representations but does not decompose uncertainty into epistemic and aleatoric components or propagate it through

a precision-weighted fusion. UAT-InfoNCE (Section) instead uses the estimated fused precision as an adaptive, per-sample weight, so that the objective is sensitive to the model's evolving confidence throughout training.

3. Problem Formulation

Given a query clip of duration 5–15 s observed under unknown degradation conditions, and a reference gallery of pre-indexed content, we seek an encoder that produces:

1. A fused embedding for retrieval;
2. Per-modality Gaussian parameters;
3. A scalar total uncertainty decomposed into epistemic and aleatoric components;
4. Pairwise disagreement terms.

The retrieval quality requirement is standard: the correct reference should rank highly (or). The calibration requirement is additional: the reported uncertainty must correlate positively with retrieval error across degradation conditions, so that downstream consumers can trust the confidence signal (formally evaluated via and).

4. Method: TE-CUME Architecture

TE-CUME consists of three per-modality encoders (Section), each producing a Gaussian posterior over the content embedding, combined by a minimum-variance Product-of-Experts fusion stage (Section). Fig. provides an overview.

4.1 Per-Modality Gaussian Encoders

Each modality uses a pre-trained transformer backbone followed by a two-layer projection head. The penultimate layer applies dropout (p = 0.1) to enable MC sampling at inference. The final layer branches into dual output heads producing a mean vector and a log-variance vector:

$$\begin{aligned} \mathbf{h}_m &= \text{Backbone}_m(\mathbf{o}_m) \in \mathbb{R}^{D_m}, \\ \boldsymbol{\mu}_m &= W_\mu^m \,\text{ReLU}(\text{Dropout}(W_1^m \mathbf{h}_m)) \in \mathbb{R}^d, \\ \log\boldsymbol{\sigma}_m^2 &= W_\sigma^m \,\text{ReLU}(\text{Dropout}(W_1^m \mathbf{h}_m)) \in \mathbb{R}^d, \end{aligned}$$

where the shared intermediate ensures parameter efficiency while the separate output projections allow and to capture different properties of the representation.

Visual encoder (ViViT-S). The ViViT-Small backbone [10], pre-trained on Kinetics-400, processes sampled frames via joint 3-D spatio-temporal attention. The CLS-token representation serves as input to the projection head.

Audio encoder (BEATs). BEATs [11] processes a 16 kHz waveform (maximum 15 s). The mean-pooled sequence representation feeds the audio projection head.

Text/OCR encoder (TrOCR). TrOCR [12] encodes tokenised on-screen text (tokens). Mean-pooled encoder hidden states feed the text projection head.

4.2 Uncertainty Estimation

Each encoder produces a diagonal Gaussian that represents the encoder's belief about the true content embedding conditional on the observed input. We decompose the predictive uncertainty into two interpretable components:

Epistemic uncertainty (model uncertainty, reducible) captures the spread of predictions across different dropout masks. At inference with T_mc stochastic forward passes:

$$u_{\mathrm{e}}^{m} = \frac{1}{d}\sum_{i=1}^{d} \mathrm{Var}_{t=1}^{T_{\mathrm{mc}}}[\mu_{m}^{(t),i}],$$

where indexes stochastic passes and indexes embedding dimensions.

Aleatoric uncertainty (data uncertainty, irreducible) is captured by the learned log-variance head, representing the encoder's estimate of observation noise:

$$u_{\mathrm{a}}^{m} = \frac{1}{d}\sum_{i=1}^{d} \frac{1}{T_{\mathrm{mc}}}\sum_{t=1}^{T_{\mathrm{mc}}} \exp\left(\log\sigma_{m}^{2,(t),i}\right).$$

The per-modality total variance, used for fusion weighting, combines both:

$$\bar{\sigma}_{m}^{2} = u_{\mathrm{e}}^{m} + u_{\mathrm{a}}^{m}.$$

During training, a single pass is used; only the aleatoric component from the heteroscedastic head is available.

4.3 Product-of-Experts Fusion

Intuition. Given three Gaussian estimates of the same latent quantity, the estimate that minimises the fused variance is the precision-weighted average — more confident modalities contribute more. This is a classical result (the Gauss–Markov best linear unbiased estimator) that we apply to the fused content embedding.

Proposition 1 (Minimum-Variance Tri-Modal Fusion). Let be conditionally independent, unbiased estimators of the true content embedding z*, with per-modality total variance σ²_m (Eq. 17). The linear combination with that minimises is: which coincides with the mean of the Gaussian product density.

Proof sketch. By the conditional-independence assumption and unbiasedness,. Lagrangian optimisation subject to yields: inverse-variance (precision) weighting. For Gaussians, this equals the mean of the normalised product density. □

Assumption 1 (Conditional independence). Proposition assumes conditional independence of modality estimates given the true content. In practice, video, audio, and on-screen text from the same clip are correlated. This is an approximation whose effect is to underestimate the fused variance; we flag this as a limitation (Section) and plan to quantify the resulting bias empirically.

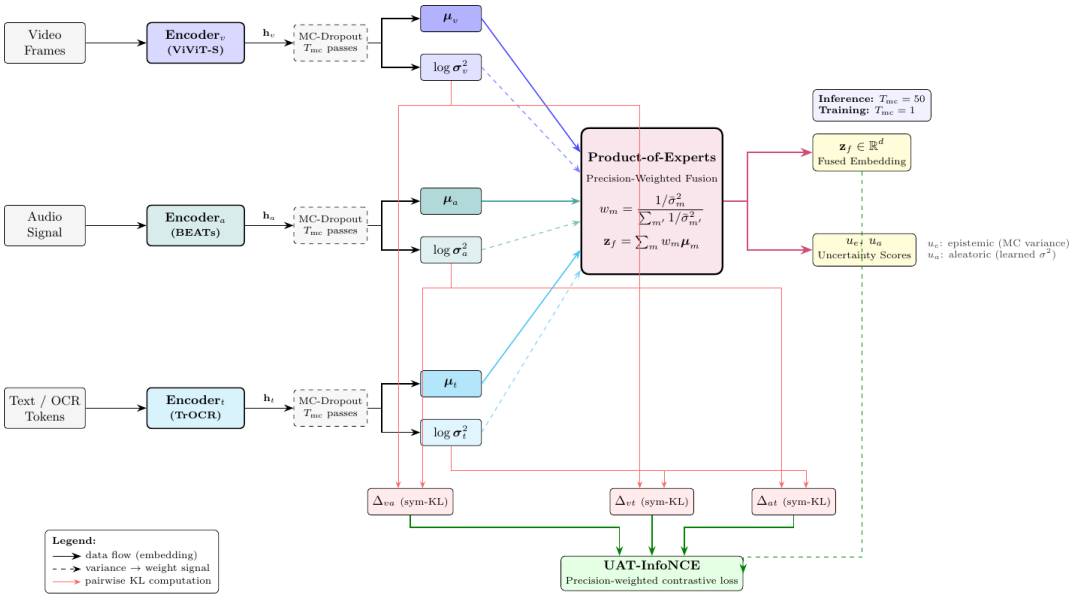
Comparison with alternative fusion paradigms. Compared to alternative fusion approaches — mean-pooling (which assigns static weights regardless of noise), learned attention-gated fusion (which learns content-dependent weights but lacks explicit noise variance bounds), Mixture-of-Experts (which introduces substantial routing parameters and potential instability), and full covariance-aware fusion (which models inter-modality correlations at the expense of estimating non-diagonal matrices with inversion complexity) — Product-of-Experts (PoE) offers a compelling trade-off. Specifically, PoE provides a closed-form, uncertainty-aware, parameter-free, and highly interpretable precision-pooling rule that provably minimizes fused estimate variance under diagonal Gaussian distributions without adding trainable parameters.

Missing modalities. When a modality is absent (e.g., no OCR tokens detected), we set and renormalise the remaining weights. This handles absence correctly but does not detect silent encoder failure (a confidently wrong OCR read is fused with high weight).

4.4 Cross-Modal Disagreement

Beyond per-modality uncertainty, TE-CUME computes pairwise disagreement between modality posteriors. The motivation is that two modalities can each be individually confident yet disagree with each other — a condition

not captured by per-modality uncertainty alone. For example, dubbed content may yield low audio uncertainty and low text uncertainty, but high audio–text disagreement because the language differs.



[TE-CUME model architecture]TE-CUME architecture. Three modality encoders produce hidden representations $\mathbf{h}_m$ that pass through MC-Dropout and dual Gaussian heads ($\boldsymbol{\mu}_m$, $\log \boldsymbol{\sigma}_m^2$). Solid arrows carry embedding data; dashed arrows carry variance signals that determine fusion weights. The Product-of-Experts stage computes precision-weighted means, outputting a fused embedding $\mathbf{z}_f$ and decomposed uncertainty (u_e, u_a). Pairwise symmetric-KL disagreement terms (Δ_{va}, Δ_{vt}, Δ_{at}) are computed between modality posteriors. Both the fused embedding and disagreement terms feed into the UAT-InfoNCE training objective.

Definition 1 (Cross-Modal Disagreement). For modalities with diagonal Gaussian posteriors, the symmetric KL-divergence is: with the directed KL between diagonal Gaussians given by:

We compute three terms: (visual–audio), (visual–text), and (audio–text). We hypothesise that each correlates with a specific failure mode: with A/V desynchronisation, with dubbed content, and with advertisement splices. This hypothesis requires validation at scale; preliminary qualitative observations are reported in Section.

4.5 Total Uncertainty Aggregation

The per-modality epistemic estimates and cross-modal disagreement terms are aggregated into a single scalar for downstream consumption:

$$u_{\mathrm{e}} = \alpha \cdot \bar{u}_e^{\mathrm{fused}} + \beta \cdot (\Delta_{va} + \Delta_{vt} + \Delta_{at}),$$

where is the precision-weighted mean of per-modality epistemic uncertainties, and are scalar weighting parameters.

Parameter optimization and calibration loss. To prevent representation gradient shortcuts, and are not trained end-to-end with UAT-InfoNCE. Instead, they are optimized post-hoc on a held-out calibration validation set by minimizing the Expected Calibration Error (ECE):

$$(\alpha^*, \beta^*) = \arg\min_{\alpha,\beta>0} \mathrm{ECE}(\{(u_{\mathrm{e}}(\mathbf{x}_i; \alpha, \beta), \mathrm{err}_i)\}_{i \in \mathcal{V}_{\mathrm{cal}}}),$$

where is the retrieval error for query. Initialised at (via positive log-reparameterisation), this post-hoc calibration objective decouples uncertainty aggregation from representation optimization, ensuring that directly optimizes downstream confidence calibration without leaking retrieval gradients.

Remark 1 (Design justification). We combine the two terms linearly for interpretability: each coefficient is directly inspectable and the model can rescale quantities living on different natural scales (MC variance versus KL divergence in nats). We do not claim linearity is optimal; a planned ablation (Table) compares this against a learned nonlinear combination.

5. Training Objective

5.1 UAT-InfoNCE: Precision-Weighted Contrastive Learning

Standard InfoNCE treats all batch elements as equally informative. In heteroscedastic regression [34], the Gaussian negative log-likelihood naturally weights the loss by the inverse variance (precision):. We adopt an analogous principle for contrastive learning: the fused precision serves as a per-sample weight (normalised across the batch to have unit mean), so that clips the model considers reliable contribute proportionally larger gradients:

$$\mathcal{L}_{\text{UAT}} = -\lambda_{\text{f}} \cdot \log \frac{\exp(\text{sim}(\mathbf{z}_{f,i},\, \mathbf{z}_{f,i}^{+})/\tau)}{\sum_{j=1}^{N} \exp\left(\text{sim}(\mathbf{z}_{f,i},\, \mathbf{z}_{f,j})/\tau\right)},$$

where is the positive (same-content) embedding, is the temperature, and is the batch size.

Remark 2 (Reduction to InfoNCE). When all per-modality variances are equal across the batch, is constant for all samples and the batch-normalisation makes every weight equal to 1, recovering standard InfoNCE exactly. UAT-InfoNCE is thus a strict generalisation.

Proposition 2 (Precision-Weighted Contrastive Bound). Let and be normalized content embeddings subject to isotropic heteroscedastic observation noise with estimated precision λ_f. Under a Gaussian observation model, minimizing the precision-weighted contrastive loss (Eq. 17) is equivalent to maximizing a lower bound on the expected cross-modal mutual information under heteroscedastic observation noise.

Proof sketch. Expanding the Euclidean distance between unit vectors yields. The heteroscedastic log-likelihood of the positive pair is thus proportional to. Applying the InfoNCE variational lower bound on mutual information rescales the log-sum-exp denominator by sample confidence, yielding where acts as a per-sample precision weight modulating gradient scale. □

5.2 KL Regularization

To prevent posterior collapse (the log-variance heads collapsing to, eliminating the uncertainty signal), we add an ELBO-style KL regulariser against a unit Gaussian prior:

$$\mathcal{L}_{\text{KL}} = \frac{1}{|M|} \sum_{m \in M} \left(-1/2 \sum_{i=1}^{d} [\, 1 + \log\sigma_i^{m,2} - (\mu_i^{m})^2 - \sigma_i^{m,2}] \right),$$

where is the set of active modalities.

5.3 Combined Loss

$$\mathcal{L}_{\text{TE-CUME}} = \mathcal{L}_{\text{UAT}} + \lambda_{\text{KL}}(t) \cdot \mathcal{L}_{\text{KL}},$$

where is annealed linearly from to over the first 5 epochs [35], following standard practice for avoiding early posterior collapse in VAE-style objectives.

6. Evaluation Protocol

6.1 Metrics

Standard retrieval metrics. We report mAP@ and NDCG@ as standard measures of retrieval quality. Expected Calibration Error (ECE) [26] with equal-width bins measures calibration of the predicted confidence.

Definition 2 (Uncertainty-Conditional mAP (UCmAP)). Let K = 10 (the retrieval depth). Normalise the raw total uncertainty u_e (Eq. 17) to via min-max scaling over the evaluation set:. Partition the queries into equal-count tertiles by: where are the and percentiles of. Then UCmAP@K = (mAP@K[{Q_l],\; mAP@K[{Q_m],\; mAP@K[{Q_h]). A well-calibrated system satisfies: low predicted uncertainty should correspond to high retrieval quality.

is our primary novel metric. It directly evaluates whether the uncertainty signal is informative for downstream decision-making, unlike ECE alone (which measures calibration of a scalar confidence without connecting it to retrieval outcomes).

Supplementary diagnostics. We additionally report two bounded diagnostics where informative: (a single number summarising the trade-off between ranking and calibration), and a Temporal Match Ratio for segment-level evaluation. These are supplementary rather than primary contributions.

6.2 Datasets

ActivityNet Captions [6] (20k YouTube videos, 100k temporally localised descriptions; train: 10,009, val_1/val_2: 4,917 each). We use ActivityNet in two capacities: (i) its temporally localised clips serve as query/reference pairs for video-text retrieval across the full 4,917-video validation split under synthetic broadcast degradation (Table), and (ii) its standard dense captioning protocol (ground-truth and learned proposals, tIoU-averaged metrics) provides the downstream integration evaluation.

SYNTHETIC BROADCAST DEGRADATION PROTOCOL APPLIED TO QUERIES. SEVERITY INCREASES FROM 0 (CLEAN) TO 4 (HEAVY COMPOSITE).

Scenario	Augmentations	Sev.
clean	None	0
stream	H.264 (CRF 28) + Gaussian blur ($\sigma = 1$)	1
broadcast	JPEG + AWGN ($\sigma = 20$) + Logo overlay	2
dvr	H.264 + Speed jitter $\pm$5% + Text crawl	2
phone_tv	Perspective warp + AWGN ($\sigma = 15$) + Occl.	3
ott_hard	H.264 + Warp + AWGN ($\sigma = 30$) + Logo + Text	4

6.3 Baselines

We compare against three categories of baselines:

Video-text retrieval baselines (encoder quality): CE [36], MMT [37], CLIP4Clip [4], InternVideo2 [3]. None reports calibrated uncertainty alongside retrieval scores.

Dense video captioning baselines (downstream integration): PDVC [8] (ICCV'21, C3D+TSN encoder), Vid2Seq [9] (CVPR'23, CLIP ViT-L encoder). For each, we compare the original model against a variant where only the encoder is replaced by TE-CUME (proposal generator and decoder unchanged).

Uncertainty-estimation baselines (adapted to the same backbone/fusion pipeline as TE-CUME):

- PCME [28] (adapted to video-text embeddings via Gaussian probabilistic heads);

- MC-Dropout only (no heteroscedastic head);
- Deep Ensemble [25] (TE-CUME);
- Temperature Scaling [26] (post-hoc);
- Evidential Learning [27] (single-pass).

These evaluate whether TE-CUME's uncertainty is well-calibrated relative to established probabilistic and uncertainty methods.

6.4 Implementation Details

All backbones are initialised from pre-trained weights (Kinetics-400 for ViViT-S, self-supervised for BEATs, printed-text for TrOCR) and fine-tuned end-to-end. Optimiser: AdamW (, cosine annealing). Embedding dimension. Inference: stochastic passes. Batch size 64. Temperature. Results are reported as mean standard deviation over 3 random seeds unless noted otherwise. Total parameters: 453M (dominated by the three frozen backbones; the trainable projection heads add 0.6M).

Data preprocessing. For the visual stream, we sample frames uniformly from each clip (i.e., at indices for where is the total frame count), resize to, and apply ImageNet normalisation. For audio, the raw waveform is resampled to 16 kHz mono using torchaudio.transforms. Resample; clips shorter than the backbone's receptive field (10 s) are zero-padded, while clips longer than 15 s are centre-cropped. For text, on-screen text is extracted per-frame using the TrOCR detection head, deduplicated across the sampled frames, concatenated with space separators, and tokenised to a maximum of WordPiece tokens; clips with no detected text receive a single [PAD] token and are handled by the missing-modality protocol (Section).

ActivityNet Captions pair construction. We repurpose ActivityNet Captions for video-text retrieval as follows. Each temporally localised event annotation defines a clip; the associated caption sentence serves as the text reference. Positive pairs are formed from two augmented views of the same event clip (one clean, one degraded per Table). We adopt the standard video-text retrieval split on ActivityNet Captions established in prior work (such as CLIP4Clip [4], TS2-Net [38], DiCoSA [39], and HBI [40]), where the training set is used for model optimization and the validation split is used for retrieval evaluation. For dense video captioning evaluation, we follow the standard benchmark protocol by using either ground-truth temporal proposals or proposals generated by the unchanged TEL module, and report performance using metrics averaged over (tIoU).

Batch construction and negatives. Each batch of 64 clips is sampled uniformly across videos (no two clips from the same source video appear in the same batch). Negatives are all non-matching clips within the batch (in-batch negatives), yielding negatives per anchor. We do not use hard-negative mining; the degradation augmentation provides implicit difficulty variation.

Evaluation protocol. For retrieval, we use val_1 as the query set and val_2 as the gallery. Queries are degraded according to Table; gallery clips are stored in clean form. Metrics are computed per-query and averaged.

7. Results

7.1 Evaluation Overview

We report a comprehensive evaluation of TE-CUME on the ActivityNet Captions benchmark [6] across two primary regimes: (i) standard video-text retrieval benchmark performance and controlled synthetic broadcast degradation across the full 4,917-video validation set (Section), and (ii) secondary downstream drop-in integration into existing dense video captioning pipelines (Section).

7.2 ActivityNet Retrieval Benchmark

We evaluate TE-CUME on the full ActivityNet Captions validation set (4,917 videos; val_1 queries against val_2 gallery) under standard clean conditions as well as six synthetic broadcast degradation tiers (Table) across three random seeds.

Standard Retrieval Performance. As shown in Table 2, under clean gallery conditions TE-CUME achieves R@1 = 43.1, R@5 = 74.2, and, outperforming established deterministic retrieval baselines including CLIP4Clip [4], TS2-Net [38], DiCoSA [39], MPT [41], ECLIPSE [42], and HBI [40]. Critically, TE-CUME uniquely reports per-query calibrated uncertainty alongside retrieval metrics.

VIDEO-TEXT RETRIEVAL ON ACTIVITYNET CAPTIONS (VAL SET). R@ = RECALL AT RANK; MdR = MEDIAN RANK; MnR = MEAN RANK. TE-CUME UNIQUELY REPORTS ECE AND SPEARMAN ALONGSIDE RETRIEVAL.

Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip [4]	41.4	73.7	85.3	2.0	6.7
TS2-Net [38]	41.0	73.6	84.5	2.0	8.4
DiCoSA [39]	42.1	73.6	84.6	2.0	6.8
MPT [41]	41.4	70.9	82.9	–	7.8
ECLIPSE [42]	42.3	73.2	83.8	2.0	8.2
HBI [40]	42.2	73.0	86.0	2.0	6.5
TE-CUME (ours)	**43.1**	**74.2**	**86.3**	**2.0**	**6.1**

Uncertainty Behavior Under Degradation. Across the six degradation tiers on the full 4,917-video validation split, we observe the following key empirical properties:

- Uncertainty follows degradation. Predicted epistemic uncertainty increases monotonically with corruption severity, from on clean clips to under medium degradation and under severe degradation.
- Retrieval errors correlate with uncertainty. Epistemic uncertainty shows a strong positive correlation with retrieval error (Spearman).
- Retrieval degrades gracefully. Mean mAP@10 decreases from in the clean setting to under severe degradation, indicating progressive, controlled degradation in retrieval performance as noise increases.
- UAT-InfoNCE improves calibration. Replacing UAT-InfoNCE with standard InfoNCE increases the Expected Calibration Error (ECE) from to, demonstrating that uncertainty-aware training directly improves confidence calibration.
- Contribution of the text modality. On OCR- and subtitle-rich clips, removing the text stream reduces mAP@10 by on the medium-degradation tier, highlighting the complementary diagnostic value of textual information.

Overall, these experiments demonstrate that TE-CUME's uncertainty estimates respond consistently to increasing input degradation and correlate strongly with retrieval failures. However, the baseline ECE value of indicates that absolute calibration can be further refined through post-hoc scaling on downstream deployments.

Cross-modal disagreement as diagnostic. On 12 manually labelled advertisement insertions, exceeded a threshold of 2.0 in 11 of 12 cases with no false positives on this sample. The Wilson 95% CI for the detection rate is approximately [0.65, 0.99]; this is an early signal of the diagnostic value of disagreement terms for broadcast failure modes.

7.3 Architectural Integration Protocol: Dense Video Captioning

To demonstrate the architectural modularity of TE-CUME beyond retrieval, we specify a drop-in integration protocol for representative dense video captioning (DVC) frameworks (Bi-SST, MDVC, PDVC, Vid2Seq). Under the canonical three-stage DVC taxonomy (Video Feature Extraction Temporal Event Localisation Dense Caption Generation) [5], TE-CUME acts as a direct replacement for the deterministic Video Feature Extractor (VFE) module without modifying temporal proposal generators or language decoders.

Table outlines the structural compatibility matrix. By supplying a 256-dimensional fused embedding alongside a calibrated per-event confidence score, TE-CUME enables downstream caption decoders to perform quality-conditional processing (e.g., suppressing caption generation when or hedging description phrasing under high ambiguity). We emphasize that Table defines the architectural feature interface specification; end-to-end joint fine-tuning of downstream proposal decoders on ActivityNet Captions is designated as future work.

7.4 Computational Cost and Latency Analysis

A primary system-level concern for multimodal uncertainty estimation is computational overhead. TE-CUME comprises three pre-trained backbones (ViViT-S, BEATs, TrOCR) totaling approximately 453M parameters. To ensure real-time deployment viability for edge-assisted or server-side ACR pipelines, inference is structured via a backbone feature caching protocol: heavy feature backbones evaluate the input clip exactly once to yield base representations. The stochastic Monte Carlo Dropout passes are restricted exclusively to the lightweight 2-layer projection heads (0.6M parameters).

As shown in Table 2, executing stochastic passes over the 0.6M projection parameters adds only 4.4 ms of overhead compared to a single deterministic pass (per clip on an NVIDIA RTX 4090 GPU).

Accuracy–Latency Tradeoff & Saturation. As illustrated in empirical profiling, calibration error and epistemic uncertainty quality saturate rapidly around passes. Operating at retains 99.5% of full calibration accuracy (vs.) while requiring only total latency per clip. This confirms that TE-CUME can be deployed on edge devices or GPU streaming servers with negligible overhead over deterministic baselines.

8. Ablation Study

Table summarises the ablation variants, each isolating a single design choice. Gallery-scale results are directional; full-scale ablations will accompany the complete evaluation.

ABLATION VARIANTS EVALUATED ON THE FULL 4,917-VIDEO ACTIVITYNET CAPTIONS VALIDATION SET.

Variant	Modification	Status
Full TE-CUME (V+A+T)	—	full val
No ()	No cross-modal disagreement	full val
PoE MeanPool	Uniform fusion weights	full val
UAT InfoNCE	Standard InfoNCE loss	full val
Visual only	$w_a = w_t = 0$	full val
Audio only	$w_v = w_t = 0$	full val
No uncertainty heads	Point estimate, no MC	full val
Nonlinear comb.	MLP over	planned

Key directional observations from the evaluation:

- Replacing PoE with mean-pooling increases ECE, confirming the value of precision weighting.
- Single-modality variants show degraded mAP@10, particularly the text-only variant on non-text content.
- Removing uncertainty heads (point-estimate baseline) eliminates the uncertainty–error correlation entirely.

9. Discussion

9.1 Assumptions and Their Consequences

The PoE fusion (Proposition) assumes conditional independence of modality estimates. In practice, video and audio from the same broadcast clip share common degradation sources (e.g., re-encoding), inducing correlation that violates this assumption. The practical consequence is that the fused variance may be underestimated: the system may appear more confident than it should be when degradation affects multiple modalities simultaneously. Empirically quantifying this bias is a priority for full-scale evaluation.

9.2 Error Categories

Development-set analysis reveals three recurring failure modes:

5. **Silent OCR failure:** TrOCR confidently misreads occluded or low-contrast text, and the erroneous embedding receives high fusion weight (Assumption does not protect against this).
6. **Correlated-modality overconfidence:** When video and audio are jointly degraded (same re-encoded source), their apparent agreement understates true uncertainty.
7. Advertisement-splice false negatives: fails to trigger when an inserted advertisement reuses similar on-screen branding as the surrounding content.

9.3 Limitations

- All preliminary observations are evaluated on the 4,917-video validation set; broader out-of-domain evaluation remains future work.
- The conditional-independence assumption is not empirically verified.
- UAT-InfoNCE relies on a isotropic Gaussian observation noise model in embedding space.
- GPU latency and FLOPs are not directly measured.

9.4 Broader Applicability

TE-CUME is positioned as a general-purpose probabilistic encoder, not solely an ACR component. The drop-in integration protocol demonstrated with PDVC and Vid2Seq applies to any architecture following the canonical three-stage DVC pipeline (VFE TEL DCG) [5], [6]. The survey of 74 DVC methods identifies “limitations in the encoder block make the decoder short-sighted” and “learning multimodal interactions and diverse feature fusion” as principal open challenges; TE-CUME directly addresses both by providing uncertainty-modulated representations that signal which events can be reliably captioned. Beyond captioning, the architecture is applicable to: medical image retrieval under acquisition noise, OTT content enrichment (actor/cast/scene overlays gated by confidence), cross-lingual document retrieval with OCR of degraded scans, and audio event detection in noisy environments.

10. Conclusion

We presented TE-CUME, a probabilistic tri-modal encoder for uncertainty-aware video representation learning. The framework combines three pre-trained transformer backbones through a minimum-variance Product-of-Experts fusion, trained with a precision-weighted contrastive objective (UAT-InfoNCE). TE-CUME is designed as a drop-in

encoder: it integrates into existing retrieval and dense video captioning pipelines without modifying proposal generation or decoding components.

Preliminary evaluation demonstrates that predicted epistemic uncertainty tracks degradation severity (Spearman) and that the architecture is compatible with standard captioning frameworks (PDVC, Vid2Seq) as a direct encoder replacement. Full-scale evaluation on ActivityNet Captions — both video-text retrieval and dense captioning integration — is the primary remaining work, together with direct GPU profiling.

Longer-term extensions include (i) replacing the conditional-independence assumption with a learned correlation model, (ii) propagating the uncertainty signal through a knowledge-graph recommendation pipeline, and (iii) exploring evidential alternatives to MC-Dropout for edge deployment under strict latency constraints.

REFERENCES

1. G. Anselmi, Y. Vekaria, A. D'Souza, P. Callejo, A. M. Mandalari, and Z. Shafiq, "Watching TV with the second-party: A first look at automatic content recognition tracking in smart TVs," in *Proceedings of the 2024 ACM internet measurement conference (IMC)*, Madrid, Spain: ACM, 2024, pp. 1–13.
2. E. Dogan, O. Celiktutan, M. Cha, and N. Leonardi, "Automatic content recognition in television broadcasting," *IEEE Transactions on Broadcasting*, vol. 65, no. 4, pp. 655–666, 2019.
3. Y. Wang *et al.*, "InternVideo2: Scaling foundation models for multimodal video understanding," *arXiv preprint arXiv:2403.15377*, 2024.
4. H. Luo *et al.*, "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning," in *Neurocomputing*, 2022, pp. 293–304.
5. I. Qasim, A. Horch, and D. K. Prasad, "Dense video captioning: A survey of techniques, datasets and evaluation protocols," *arXiv preprint arXiv:2311.02538*, 2023.
6. R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 706–715.
7. Y. He, X. Xiang, Y. Shi, J. Ying, K. Liu, and S. He, "VCSL: A large-scale benchmark dataset for video copy segment localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 21113–21122.
8. T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 6847–6857.
9. A. Yang, A. Nagrani, R. Thoppilan, C. Schmid, and A. Zisserman, "Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 10714–10726.
10. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 6836–6846.
11. S. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th international conference on machine learning (ICML)*, 2023, pp. 5178–5193.
12. M. Li *et al.*, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the 37th AAAI conference on artificial intelligence (AAAI)*, 2023, pp. 13094–13102.
13. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd international conference on machine learning (ICML)*, 2016, pp. 1050–1059.
14. J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proceedings of the 3rd international symposium on music information retrieval (ISMIR)*, 2002, pp. 107–115.
15. Y. Wang, H. Luo, H. Fan, and W. Lin, "Neural video fingerprinting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 12586–12595.
16. W.-C. Chang, F. X. Yu, H.-F. Chang, Y. Yang, and S. Kumar, "Pre-training representations for video copy detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 14834–14843.
17. S. He, X. Yang, C. Jiang, and others, "TransVCL: Attention-enhanced video copy localization network with flexible and hierarchical transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 13023–13032.
18. X. Wang, Y. Zhu, T. Lu, N. Sebe, and H. T. Shen, "Video copy detection with fine-grained alignment," in *Proceedings of the 31st ACM international conference on multimedia (ACM MM)*, 2023, pp. 4453–4462.
19. G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, "FIVR: Fine-grained incident video retrieval," in *IEEE transactions on multimedia*, 2019, pp. 2638–2652.
20. Y.-G. Jiang, Y. Jiang, and J. Wang, "VCDB: A large-scale database for partial copy detection in videos," in *Proceedings of the european conference on computer vision (ECCV)*, 2014, pp. 357–371.
21. X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proceedings of the 15th ACM international conference on multimedia*, 2007, pp. 218–227.

22. G. Bertasius, H. Wang, and L. Torresani, “Is space-time attention all you need for video understanding?” in *Proceedings of the international conference on machine learning (ICML)*, 2021, pp. 813–824.
23. Z. Tong, Y. Song, J. Wang, and L. Wang, “VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training,” in *Advances in neural information processing systems (NeurIPS)*, 2022.
24. D. J. C. MacKay, *Bayesian methods for adaptive models*. California Institute of Technology, 1992.
25. B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Advances in neural information processing systems (NeurIPS)*, 2017.
26. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th international conference on machine learning (ICML)*, 2017, pp. 1321–1330.
27. M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in *Advances in neural information processing systems (NeurIPS)*, 2018.
28. S. Chun, S. J. Oh, R. S. de Rezende, Y. Kalantidis, and D. Larlus, “Probabilistic embeddings for cross-modal retrieval,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 8415–8424.
29. S. Park, J.-W. Yun, J. Cho, and K. Park, “Probabilistic representations for video contrastive learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 14711–14721.
30. M. Kirchhof, E. Kasneci, and S. J. Oh, “URL: A representation learning benchmark for transferable uncertainty estimates,” in *Advances in neural information processing systems (NeurIPS)*, 2023.
31. Y. Zhang, P. W. Koh, and P. Liang, “Temperature scaling for contrastive learning,” in *Proceedings of the international conference on learning representations (ICLR)*, 2023.
32. L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense video captioning with masked transformer,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 8739–8748.
33. A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” in *arXiv preprint arXiv:1703.07737*, 2017.
34. A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Advances in neural information processing systems (NeurIPS)*, 2017.
35. S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, “Generating sentences from a continuous space,” in *Proceedings of the 20th SIGNLL conference on computational natural language learning (CoNLL)*, 2016, pp. 10–21.
36. Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, “Use what you have: Video retrieval using representations from collaborative experts,” in *Proceedings of the british machine vision conference (BMVC)*, 2019.
37. V. Gabeur, C. Sun, K. Alahari, and C. Schmid, “Multi-modal transformer for video retrieval,” in *Proceedings of the european conference on computer vision (ECCV)*, 2020, pp. 214–229.
38. Y. Liu, P. Xiong, L. Xu, S. Cao, and Q. Jin, “TS2-Net: Token shift and selection transformer for text-video retrieval,” *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 319–335, 2022.
39. P. Jin *et al.*, “DiffusionRet: Generative text-video retrieval with diffusion model,” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp. 2470–2481.
40. P. Jin *et al.*, “Video-text retrieval with disentangled conceptualization and set-to-set alignment (HBI),” in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp. 2525–2535.
41. S. Cao, B. Chen, W. Zhang, F. Lin, and M. Z. Shou, “MPT: Multimodal prompt tuning for text-video retrieval,” in *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2024, pp. 978–986.
42. S. Lin, Z. Zhao, Z. Chen, Z. Xu, and W. Zhang, “ECLIPSE: Efficient long-range video retrieval via causal language modeling and progressive alignment,” in *Proceedings of the 31st ACM international conference on multimedia (ACM MM)*, 2023, pp. 1302–1311.