Qwen-NEWS: Video Summarization Model for News-Style Reports

Che Tian [1] **Zhiyuan Li** [1] **Lei Tang** [1] **Hongrui Song** [1]

Abstract

Video has become a dominant medium for news, but it is inherently linear and costly to consume. We address this inefficiency by proposing Qwen-NEWS, a multimodal video-to-article framework that converts news videos into structured textual reports with *Title*, *Highlights*, and *Article* sections. Our pipeline uses Whisper to transcribe audio, BLIP to caption adaptively selected keyframes, and a LoRA-adapted Qwen-2.5-32B-Instruct model to synthesize the final article from a unified multimodal context. We evaluate on a curated subset of the MM-AVS benchmark and compare against strong prompt-only baselines, including FLAN-T5 XXL zero-shot and Qwen-32B zero-/few-shot. Qwen-NEWS achieves higher BERTScore, ROUGE, and METEOR, and LoRA fine-tuning reduces validation negative log-likelihood by roughly 14% relative to the frozen backbone, indicating effective domain adaptation. Qualitative analysis shows that the model produces coherent, fact-dense articles when audio and visual evidence are complementary, but can hallucinate plausible yet unsupported scene details in visually homogeneous clips. These results suggest that structured news-style video summarization is feasible with current LLMs, while motivating future work on tighter audio–visual alignment and grounding-aware evaluation.

1. Introduction

Problem Statement and Motivation In the modern information age, video has become a primary medium for news dissemination. While this format provides multi-modal context, it presents a significant bottleneck for information processing. Consuming video news is inherently linear and time-consuming. This means that researchers and the public must spend hours watching footage to extract key information, a process that is inefficient and scales poorly.

Furthermore, a simple audio transcription, while useful, only captures one dimension of the story. It can sometimes miss critical non-verbal information, such as the visual context of an event or the identification of a key figure. This "modality gap" results in an incomplete and often decontextualized understanding.

Our project aims to solve this problem by developing an AI assistant that leverages the multimodal nature of video. We propose a system that automatically converts a news video into a structured, textual news article. This output is not just a summary; it is a new, coherent artifact that is dense, searchable, and synthesizes both the auditory and visual channels. This system provides immediate value by enabling rapid comprehension and facilitating downstream tasks, such as fact-checking and topic analysis.

Role of Deep Learning and Generative Models Addressing this multi-modal challenge is intractable with classical computational methods. The project's success fundamentally relies on the advanced capabilities of large-scale deep learning and generative models, which play three distinct roles.

1. A robust automatic speech recognition (ASR) model is required for the perception task of converting the continuous, often noisy, audio stream into a clean textual transcript.

2. A vision-language model (VLM) is essential to bridge the "modality gap" by performing visual perception, interpreting key video frames to extract semantic context that is absent from the audio.

3. A generative text-to-text model, likely an instruction-tuned transformer, serves as the central synthesis engine. Its role is not merely summarization, but the complex task of reasoning over, fusing, and re-formatting two distinct and unaligned textual inputs (one verbose, one sparse). This model must learn to follow structural commands and generate a coherent, novel article. This project therefore explores the synergy of specialized

[1]STAT 453 Course Project. Correspondence to: Benjamin Lengerich <lengerich@wisc.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

generative models to achieve a complete, multi-step transformation of information.

2. Related Work

Video summarization has progressed from visual-only keyframe selection to multimodal pipelines that fuse audio, text, and vision. A recent survey formalizes summarization as (i) selecting salient snippets and (ii) composing a compact narrative, and emphasizes *input-driven* conditioning (text prompts, reference clips, full video, or audio). It also notes the absence of a clear state of the art for *audio-only* summarization and the persistent challenge of synchronizing speech with visuals (Meena et al., 2023). The survey further reviews common datasets (SumMe, TVSum) and metrics (F1, coverage, diversity), underscoring open issues in redundancy control and temporal alignment.

On the speech side, Whisper shows that large-scale, weakly supervised training yields robust, zero-shot ASR across domains and languages without task-specific fine-tuning (Radford et al., 2022). Its multilingual, multitask objective and timestamped decoding are particularly useful for aligning spoken claims with visual events in noisy, in-the-wild videos.

For text generation, the Text-to-Text Transfer Transformer (T5) unified diverse NLP tasks under a single sequence-to-sequence interface and established strong encoder–decoder baselines for summarization (Raffel et al., 2020). Pretraining via span corruption on C4 and prompt-style task prefixes enables controllable outputs while preserving factual conditioning from evidence.

Instruction tuning (e.g., FLAN) further demonstrated that fine-tuning on natural-language instructions substantially improves zero/few-shot generalization to unseen tasks (Chung et al., 2024). Template variation and broad task mixtures reduce prompt sensitivity, improving adherence to requested formats such as headlines, leads, and citations.

On the vision–language side, BLIP introduced bootstrapped pre-training that filters and improves web image–text pairs, supporting both bidirectional (understanding) and causal (generation) modes with strong captioning and retrieval performance (Li et al., 2022). Its ViT-based encoder and multimodal transformer yield frame captions that surface entities and quantities, aiding downstream grounding with ASR.

Taken together, these advances provide robust ASR, instruction-following generation, and high-quality vision–language representations, yet leave open an *end-to-end, timeline-grounded, citation-aware* approach tailored to news writing. Our project targets this gap by aligning transcripts and frame captions at the evidence level, composing a controllable news-style report, and attaching quantitative checks for structure, grounding, and temporal alignment.

3. Methods

3.1. Overview of the Proposed Framework

We propose a comprehensive multi-modal framework designed to transform raw video inputs into structured news articles. As illustrated in Figure 1, our approach decomposes the input video into two parallel processing streams: an *audio stream* for extracting verbal content and a *visual stream* for capturing scene semantics. These modalities are subsequently integrated via a multi-modal fusion module, which serves as the context for a Large Language Model (LLM) to generate the final article.

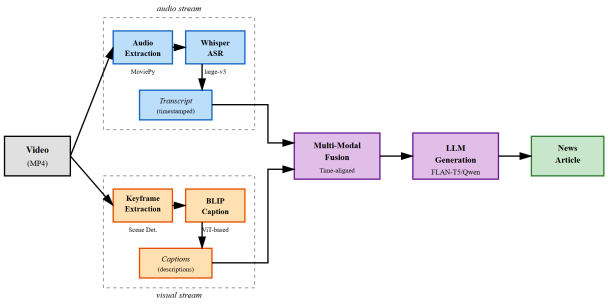

Figure 1. Schematic illustration of the proposed pipeline.

Formally, given an input video $\mathcal{V}$, our goal is to generate a textual document $\mathcal{Y}$. The pipeline consists of three main stages:

- **Modality Extraction:** We first separate the video into audio signals and visual frames. The audio is processed to yield a transcript sequence $T = \{t_1, t_2, \ldots, t_n\}$, while the visual frames are converted into a sequence of descriptive captions $C = \{c_1, c_2, \ldots, c_m\}$.
- **Context Aggregation:** We adopt a direct concatenation strategy to integrate the modalities. The transcript sequence T and the set of visual descriptions C are merged sequentially to form a unified global context $\mathcal{Z}$.
- **LLM Generation:** $\mathcal{Z}$ is fed into a fine-tuned Large Language Model to synthesize the news article $\mathcal{Y}$.

3.2. Modality Extraction

3.2.1. Audio Stream Processing

The primary goal of the audio stream is to recover the verbal narrative and spoken content embedded in the video. The processing consists of two stages: acoustic signal extraction and automatic speech recognition.

Acoustic Signal Extraction. Given a raw video input $v \in \mathcal{V}$, we first isolate the continuous audio track to eliminate visual redundancy and focus on linguistic information. Formally, we define a demultiplexing function $\Phi_{\text{audio}}(\cdot)$ that separates the audio channels from the video container:

$$a = \Phi_{\text{audio}}(v), \tag{1}$$

where a represents the extracted acoustic waveform. In our implementation, this step involves decoding the audio stream from the MP4 container and re-encoding it into a standardized MP3 format (44.1 kHz sample rate) to ensure compatibility with the downstream transcription module. This extraction process acts as a filter, removing visual noise while preserving the temporal structure of the speech.

Robust Speech Transcription. Upon extracting the raw waveform a, we aim to generate a high-fidelity textual transcript $\mathcal{T}$. A naive application of ASR models often yields suboptimal results due to ambient noise and model hallucinations. To mitigate these issues, we propose a multi-stage transcription pipeline.

First, to filter out non-speech segments (e.g., background music, wind noise), we apply a **Voice Activity Detection (VAD)** pre-filter. We utilize the Silero VAD engine, retaining only those audio segments where the speech probability exceeds a strict threshold τ_{vad}. This ensures that the downstream model focuses solely on clear human articulation.

Subsequently, the filtered audio is processed by the `whisper-large-v3`, a Transformer-based sequence-to-sequence estimator, to produce the initial candidate text $\hat{\mathcal{T}}$.

$$\hat{\mathcal{T}} = \text{Whisper}(a | \text{VAD}(a) > \tau_{\text{vad}}). \tag{2}$$

Finally, we implement a **Hallucination Suppression** mechanism. Neural ASR models are known to generate repetitive artifacts or training-data watermarks (e.g., "Thanks for watching", "Subtitles by...") when the audio is silent or ambiguous. We employ a linguistic validity filter $\mathcal{F}_{\text{clean}}$ that detects and removes these specific hallucinatory phrases and non-English gibberish, yielding the final transcript $T = \mathcal{F}_{\text{clean}}(\hat{\mathcal{T}})$.

3.2.2. Visual Stream Processing

Complementary to the audio stream, visual cues provide essential grounding for scene understanding and object identification. However, processing high-dimensional video data directly is computationally difficult and replete with redundancy. To address this, we propose a two-stage visual pipeline: adaptive keyframe extraction followed by semantic captioning.

Adaptive Keyframe Extraction. To condense the video into a discrete sequence of informative images, we employ an **Adaptive Shot Boundary Detection (SBD)** algorithm. Unlike fixed-interval sampling (e.g., extracting 1 frame every second) which often captures blurry transitions or redundant static frames, our approach dynamically segments the video based on visual content changes.

Let $\mathcal{V} = \{f_1, f_2, \ldots, f_T\}$ represent the raw video frames. We calculate the frame-to-frame difference score $\Delta(f_t, f_{t+1})$ based on pixel intensity changes in the HSV color space. A new scene segment S_k is identified whenever Δ exceeds a sensitivity threshold τ_{scene}. This yields a set of temporal segments $S = \{S_1, S_2, \ldots, S_K\}$.

For each segment S_k, we select a single representative keyframe v_k to encapsulate the visual semantics. To mitigate motion blur common at shot boundaries, we sample the frame at the temporal centroid of the segment:

$$v_k = \mathcal{V}\left(\frac{t_{\text{start}}^{(k)} + t_{\text{end}}^{(k)}}{2}\right). \tag{3}$$

In scenarios with minimal visual variance (e.g., static slides), we implement a fallback mechanism that performs global temporal pooling to extract a median frame. The final output is a curated sequence of keyframes $\mathcal{K} = \{v_1, v_2, \ldots, v_K\}$.

Semantic Caption Generation. To bridge the modality gap between low-level visual features and high-level textual semantics, we employ a pre-trained Vision-Language Model (VLM) to interpret the selected keyframes. Specifically, we utilize the `blip-image-captioning-large` architecture with the ViT-Large backbone.

For each keyframe $v_k \in \mathcal{K}$, the model generates a natural language description c_k. The generation process is formulated as finding the token sequence that maximizes the conditional likelihood:

$$c_k = \underset{c}{\text{argmax}}\, P(c \mid v_k; \theta_{\text{BLIP}}). \tag{4}$$

To ensure high-fidelity descriptions, we employ a **Beam Search** decoding strategy to enforce concise descriptions, preventing verbose hallucinations. The final visual representation is a sequence of chronological captions $C = \{c_1, c_2, \ldots, c_K\}$, which is now textually compatible with the audio transcript.

3.3. Context Aggregation

The aggregation process consists of two steps: heterogeneous content linearization and paradigm-specific prompt construction.

3.3.1. Heterogeneous Content Linearization

Given the extracted audio transcript T (Section 3.2.1) and the sequence of visual captions C (Section 3.2.2), we first se-

rialize these modalities into a unified textual format. To preserve the source distinction, we introduce modality-specific separators. The linearized context $\mathcal{X}_{\text{ctx}}$ is constructed as:

$$\begin{aligned}\mathcal{X}_{\text{ctx}} = &\texttt{[Visual Start]} \oplus C \oplus \texttt{[Visual End]} \\ &\oplus \texttt{[Audio Start]} \oplus T \oplus \texttt{[Audio End]},\end{aligned} \tag{5}$$

where $\oplus$ denotes string concatenation. This serialization ensures that the LLM can explicitly distinguish between what was *seen* (visual semantics) and what was *heard* (narrative content).

3.3.2. Prompting Paradigms

We investigate three distinct prompting paradigms, corresponding to the varying levels of supervision in our experimental setup:

Zero-Shot Instruction Format. For our baseline evaluators (e.g., FLAN-T5, Qwen-ZeroShot), we construct the input $\mathcal{Z}_{\text{zero}}$ by prepending a system-level role definition $\mathcal{I}_{\text{sys}}$ ("You are a professional news editor...") and a strict formatting constraint $\mathcal{I}_{\text{fmt}}$ ("Structure the report with Title, Highlights, and Article"):

$$\mathcal{Z}_{\text{zero}} = \mathcal{I}_{\text{sys}} \oplus \mathcal{I}_{\text{fmt}} \oplus \mathcal{X}_{\text{ctx}}. \tag{6}$$

In-Context Learning (Few-Shot). To guide the model's style and length compliance without weight updates, we also employ few-shot prompting. Let $\mathcal{D}_{\text{train}}$ be the training corpus. We dynamically retrieve k demonstration pairs $\{(x_i, y_i)\}_{i=1}^{k} \sim \mathcal{D}_{\text{train}}$ to form the context history. The input $\mathcal{Z}_{\text{few}}$ is formulated as:

$$\begin{aligned}\mathcal{Z}_{\text{few}} = &\mathcal{I}_{\text{sys}} \\ &\oplus \underbrace{\sum_{i=1}^{k} (\texttt{User}: x_i \oplus \texttt{Assistant}: y_i)}_{\text{Demonstration Context}} \\ &\oplus (\texttt{User}: \mathcal{X}_{\text{ctx}}).\end{aligned} \tag{7}$$

Instruction Tuning Format (SFT). For our proposed fine-tuned model (Qwen-SFT), we adopt the Alpaca-style template to align with the supervised fine-tuning stage. The context is structured into explicit XML-like sections to maximize instruction adherence:

$$\begin{aligned}\mathcal{Z}_{\text{sft}} = &\texttt{\#\#\# Instruction:} \oplus \mathcal{I}_{\text{sys}} \\ &\oplus \texttt{\#\#\# Input:} \oplus \mathcal{X}_{\text{ctx}} \\ &\oplus \texttt{\#\#\# Response:}.\end{aligned} \tag{8}$$

This structured prompt explicitly signals the model to generate the completion starting from the response token.

3.4. LLM Generation

Given the aggregated context $\mathcal{Z}$ (Section 3.3), the final stage aims to generate the structured news report $\mathcal{Y}$. We investigate two distinct backbone architectures and propose a parameter-efficient fine-tuning strategy to adapt general-purpose LLMs to this multimodal summarization task.

3.4.1. Architectures and Quantization

To evaluate the efficacy of our framework across different model families, we employ two representative architectures:

Encoder-Decoder (FLAN-T5). We utilize `FLAN-T5 XXL (11B)` as our baseline. As a sequence-to-sequence model, it processes the input context bidirectionally, making it inherently robust for summarization tasks.

Decoder-Only (Qwen). We adopt the `Qwen-2.5-32B-Instruct`, a SOTA causal language model. To mitigate the substantial memory footprint of the 32B parameter model, we employ **4-bit Normal Float (NF4) quantization**. This technique compresses the model weights while preserving zero-point precision, allowing for efficient inference on consumer-grade hardware without significant degradation in generation quality.

3.4.2. Parameter-Efficient Fine-Tuning (PEFT)

Baseline general language models often fail to strictly adhere to the complex formatting requirements of news reporting. To address this, we implement a Supervised Fine-Tuning (SFT) stage using Low-Rank Adaptation (LoRA).

Instead of updating all dense parameters Θ, LoRA freezes the pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ and injects trainable rank-decomposition matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, where the rank $r \ll \min(d, k)$. The forward pass for a target layer $h = W_0 x$ is modified as:

$$h = W_0 x + \frac{\alpha}{r} BAx, \tag{9}$$

where α is a scaling factor. In our specific implementation for Qwen-32B, we apply LoRA adapters to all linear projection layers within the attention mechanism and the feed-forward networks. The optimization objective is to minimize the negative log-likelihood of the target news report $\mathcal{Y}$ given the context $\mathcal{Z}_{\text{sft}}$:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|Y|} \log P(y_t \mid y_{<t}, \mathcal{Z}_{\text{sft}}; \Theta_{\text{LoRA}}). \tag{10}$$

3.4.3. Decoding Strategies

During inference, we tailor the decoding strategy. For the FLAN-T5 baseline, we employ **Beam Search** (width $b = 5$) to explore the hypothesis space and select the sequence with

the highest cumulative probability. Conversely, for the open-ended generation capabilities of Qwen (both baseline and fine-tuned variants), we utilize **Nucleus Sampling** (Top-p sampling) with temperature scaling. This stochastic decoding introduces diversity and reduces repetition, generating more naturalistic news narratives.

4. Experiments

4.1. Dataset Setup

4.1.1. Source Dataset: MM-AVS

We construct our experimental benchmark based on the **MM-AVS dataset** (Fu et al., 2021). Derived from CNN and Daily Mail, MM-AVS represents a comprehensive collection designed for multimodal summarization tasks.

MM-AVS ensures that each news story is associated with:

- **Textual Content:** Full-length news articles, editor-written summaries (highlights), and concise titles.
- **Visual Content:** Raw video footage and associated keyframe images with captions.
- **Acoustic Content:** Complete audio tracks extracted from the videos, paired with transcripts.

4.1.2. Data Curation and Standardization

To transform the raw corpus into a reliable benchmark, we strictly isolated the raw video footage and the editorial ground truth (Article, Highlights, and Title), discarding all pre-existing metadata (e.g., legacy transcripts or captions) to ensure that all multimodal inputs are dynamically generated by our own method.

The curation involves three filtration stages:

- **Integrity and Asset Recovery:** We performed an audit to resolve data fragmentation. Raw video streams, originally distributed as fragmented `.ts` files, were concatenated into unified MP4 containers using `ffmpeg`. Samples that lacked essential ground-truth components were pruned.
- **Lexical De-duplication:** To prevent data leakage, we implemented a de-duplication algorithm based on **normalized transcript fingerprints**. We computed a unique lexical signature for each sample. Entries sharing identical narrative content were removed, ensuring that each entry in our benchmark represents a unique event.
- **Standardization and Splitting:** The surviving valid samples were unified into a standardized JSON structure, aggregating the visual descriptions and audio transcripts. We adopted a **fixed hold-out strategy** for evaluation, designating 10 diverse samples for the Test set. The remaining data was partitioned into Training (90%) and Validation (10%) sets. This resulted in a final curated subset of **529** high-quality samples.

Table 1. Statistics of the curated dataset. "Avg. Text Len" denotes the average token count with BERT Base Tokenizer.

Split	Samples	Avg. Video Len	Avg. Text Len
Train	468	129.74s	588.31
Val	51	141.81s	721.47
Test	10	140.19s	616.30

4.2. Models and Baselines

To assess the effectiveness of our multimodal pipeline and supervised fine-tuning strategy, we compare our proposed model against three strong instruction-following LLM configurations. All models receive the same linearized multimodal context $\mathcal{X}_{\text{ctx}}$ (Equation (5)), consisting of BLIP captions and Whisper transcripts, and differ only in how the backbone LLM is adapted and prompted.

Qwen-NEWS Our primary system is the **Qwen-NEWS** model, built on `Qwen-2.5-32B-Instruct`. We apply parameter-efficient supervised fine-tuning using LoRA adapters (Section 3.4.2) on the training portion of our curated MM-AVS subset. The model is trained with the instruction-tuning format $\mathcal{Z}_{\text{sft}}$ (Equation (8)), explicitly conditioning on the multimodal context and the desired news-style structure (Title, Highlights, Article). This configuration serves as the main point of comparison in all subsequent experiments.

Qwen-2.5-32B Zero-shot The first baseline keeps the same backbone, `Qwen-2.5-32B-Instruct`, completely *frozen*. At inference time it is prompted only with the zero-shot instruction template $\mathcal{Z}_{\text{zero}}$ that describes the editing role and required output format, but the weights are not updated on MM-AVS. This setting measures how far a large, general-purpose instruction-tuned model can go on our task without any task-specific supervision.

Qwen-2.5-32B Few-shot The second baseline evaluates in-context learning with the same frozen backbone. We prepend k demonstration pairs sampled from the training set to form the few-shot prompt $\mathcal{Z}_{\text{few}}$ (Equation (7)), where each demonstration consists of a multimodal context and its corresponding ground-truth news report. This setting tests whether providing exemplars at inference time can recover the desired structure and grounding without gradient updates.

FLAN-T5 Zero-shot Finally, we include `FLAN-T5 XXL (11B)` as an encoder–decoder baseline. FLAN-T5 is prompted with the same zero-shot instruction and the same linearized multimodal context, but no task-specific fine-tuning is performed. This configuration represents a strong, widely used sequence-to-sequence summarization baseline and allows us to examine whether our gains are specific to Qwen or hold across model families.

Together, these three baselines isolate the contributions of (i) supervised fine-tuning versus pure prompting, and (ii) model architecture, while controlling for the multimodal input and instruction format.

4.3. Training Details and Evaluation Protocol

For Qwen-NEWS, we fine-tune `Qwen-2.5-32B-Instruct` with LoRA adapters as described in Section 3.4.2. The 32B backbone is loaded with 4-bit NF4 quantization on a single A100 GPU, and only the low-rank adapter parameters are updated. We optimize the token-level cross-entropy loss with AdamW, using a learning rate of $1e{-}4$, global batch size 1, and a linear warmup followed by cosine decay. Training is run for up to three epochs on the 468-sample training split; after every ten optimization steps we evaluate on the 51-sample validation split. The validation loss decreases smoothly from roughly 2.08 at the start of training to around 1.76 by the final epoch, indicating stable convergence without obvious overfitting. All reported Qwen-NEWS results use the checkpoint with the lowest validation loss.

For the three baseline systems (Qwen zero-shot, Qwen few-shot, and FLAN-T5 zero-shot), the backbone parameters are kept frozen. We apply the prompting paradigms introduced in Section 3.3—zero-shot instruction prompting, few-shot in-context prompting, and encoder–decoder zero-shot prompting, respectively—without any gradient updates on MM-AVS. All models share the same multimodal context $\mathcal{X}_{\text{ctx}}$ and the decoding strategies described in Section 3.4, ensuring a fair comparison.

We evaluate all systems on the 10-story test split using BERTScore (Precision, Recall, and F1), ROUGE-1/2/L, and METEOR computed over the Article section. BERTScore F1 measures semantic similarity between generated and reference articles in a contextual embedding space, ROUGE captures lexical overlap at the unigram, bigram, and longest-common-subsequence levels, and METEOR balances adequacy and fluency with synonym matching and fragmentation penalties. For each metric, we compute the mean and standard deviation across the 10 test stories to quantify both central performance and variability; the aggregate results are reported in Section 4.4, and representative successes and failures are analyzed qualitatively in Section 4.6.

Table 2. Quantitative results on the 10-sample test set. We report BERTScore F1, METEOR, and ROUGE scores for all model variants.

Model	BERT-F1	METEOR	R-1	R-2	R-L
FLAN-T5 (ZS)	0.8104	0.0291	0.0784	0.0234	0.0520
Qwen-32B (ZS)	0.8212	0.1353	0.2523	0.0567	0.1295
Qwen-32B (FS)	0.8194	0.1245	0.2377	0.0463	0.1203
Qwen-NEWS (SFT)	**0.8281**	**0.1562**	**0.2984**	**0.0735**	**0.1403**

4.4. Main Quantitative Results

4.4.1. Overall performance of Qwen-NEWS

On the held-out test set, Qwen-NEWS attains a BERTScore F1 of 0.8281 with precision 0.8363 and recall 0.8202, indicating strong semantic alignment between the generated articles and the human-written ground truth. The model achieves a METEOR score of 0.1562, reflecting adequate content coverage and reasonably fluent surface forms. In terms of lexical overlap, Qwen-NEWS obtains ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.2984, 0.0735, and 0.1403, respectively, which are competitive for an abstractive generation task with long, multi-paragraph outputs.

4.4.2. Comparison with prompt-only baselines

Table 2 compares Qwen-NEWS against three strong prompt-only baselines that all operate on the same multimodal inputs. FLAN-T5 XXL zero-shot underperforms substantially, with very low ROUGE scores (e.g., ROUGE-1 of 0.0784) and a METEOR score of only 0.0291; manual inspection shows that it often ignores the requested structure and generates short, generic text instead of full articles. Switching to the decoder-only Qwen-2.5-32B backbone yields a large improvement even without task-specific supervision: the zero-shot and few-shot variants reach BERTScore F1 around 0.82 and METEOR between 0.12 and 0.14, with ROUGE-1 in the 0.24–0.25 range. However, these prompt-based configurations still exhibit inconsistent Title/Highlights/Article structure and occasional hallucinations.

Supervised fine-tuning with LoRA further improves performance across all metrics. Relative to the best prompt-only baseline (Qwen zero-shot), Qwen-NEWS increases BERTScore F1 by about 0.7 points and METEOR by roughly 0.02, while raising ROUGE-1 from 0.2523 to 0.2984 and ROUGE-2 from 0.0567 to 0.0735. These gains indicate that lightweight adaptation is sufficient to make the model more faithful to the multimodal evidence and better aligned with the stylistic and structural requirements of news-style reporting.

4.5. Training Diagnostic: Impact of LoRA Fine-Tuning

To better understand how LoRA adaptation changes the behavior of the underlying `Qwen-2.5-32B-Instruct`

backbone, we conduct a small-scale training diagnostic that compares the model *before* and *after* LoRA fine-tuning. Both configurations share the same architecture, tokenizer, and multimodal inputs described in Section 3; the only difference is whether rank-16 LoRA adapters are activated and trained on MM-AVS.

For the *pre-adaptation* configuration, we evaluate the frozen instruction-tuned Qwen-2.5-32B on the training and validation splits and record the average negative log-likelihood loss. For the *LoRA-adapted* configuration (Qwen-NEWS), we report the corresponding losses at the selected checkpoint with the lowest validation loss (Section 4.3).

Table 3. Effect of LoRA fine-tuning on negative log-likelihood loss. The two columns correspond to the same backbone before and after LoRA adaptation.

Metric	Pre-adaptation	Qwen-NEWS
Training Loss	2.076	1.697
Validation Loss	2.047	1.757
Relative Reduction (val)	—	-14.2%

As shown in Table 3, enabling LoRA adaptation reduces the validation loss from 2.047 to 1.757, a relative decrease of roughly 14.2%, with a comparable drop on the training set (from 2.076 to 1.697). Since all other components are held fixed, this reduction can be attributed purely to the additional capacity provided by the low-rank adapters and their exposure to news-style supervision. Together with the metric gains reported in Table 2, this diagnostic suggests that LoRA fine-tuning allows the model to internalize news-specific patterns and structural constraints that are not fully captured by the generic instruction-tuned checkpoint.

4.6. Qualitative Analysis

Beyond aggregate metrics, we conduct a qualitative inspection of Qwen-NEWS outputs on the 10-story test set to better understand how the model uses multimodal evidence in practice. Overall, we observe that when both the transcript and the keyframe captions contain rich, complementary information, Qwen-NEWS produces articles that closely mirror human-written stories in both structure and content. For instance, in clips covering natural disasters or international conflicts, the model typically introduces the main event in the Title and first paragraph, then uses the Highlights to surface key numbers (casualty counts, dates, locations), and finally weaves together quotes from the transcript with scene descriptions from BLIP captions in the body. In these cases, entity mentions and numerical facts are well aligned with the transcript, and visual details such as "flooded streets" or "damaged buildings" can be traced directly to specific keyframe captions.

However, a recurring failure mode emerges in videos where the visual stream is low-information or highly homogeneous. Typical examples include studio segments in which a single anchor reads the news in front of an unchanging background or press briefings where a politician stands at a podium for the entire clip. In such cases, the BLIP captions tend to be short and repetitive (e.g., "a man in a suit speaking at a podium"), adding little beyond what is already implied by the transcript. When exposed to this kind of degenerate visual input, Qwen-NEWS sometimes "overcompensates" by introducing richer scene descriptions than the evidence supports. We observe summaries that mention a "crowd in the room," a "tense atmosphere," or specific venues such as "the White House briefing room" even when the video shows only a tight shot of the speaker and the transcript does not name the location. In a few instances, the model also fabricates brief cutaways (e.g., "showing images of damaged buildings" or "cutting to footage of protesters") that follow common news tropes but are absent from both the transcript and the keyframes.

These hallucinations are subtle enough that automatic metrics, which primarily reward lexical and semantic overlap with the reference article, remain relatively high, yet from a grounding perspective they violate our design goal that each visual claim should be traceable to either the transcript or the keyframe captions. Qualitative analysis thus reveals a gap between "looking good" under text-based metrics and being strictly evidence-aligned in a multimodal sense. We return to this limitation and outline potential mitigation strategies in our discussion of future work in Section 5.

5. Discussion

5.1. Limitations

Although Qwen-NEWS performs well on MM-AVS, several limitations remain. First, our adaptive keyframe selection combined with independent captioning can fail in visually homogeneous clips: as shown in Section 4.6, when the visual stream collapses to a few generic captions, the model occasionally hallucinates crowd reactions or cutaway shots that are not present in either the transcript or the frames.

Second, our evaluation is dominated by text-only automatic metrics. BERTScore and ROUGE (Section 4.4) capture semantic and lexical overlap with the reference article but only indirectly reflect multimodal grounding: a hallucinated description can still receive a high score if it resembles the human write-up. The loss-based diagnostic in Section 4.5 likewise confirms better likelihood but does not distinguish faithful from embellished explanations.

Finally, our curated MM-AVS subset is small and stylistically narrow: 529 English-language broadcast stories with a specific three-part structure. The 10-story test split limits its statistical power, and it is unclear how well our model

would generalize to local news, long-form documentaries, or non-Western outlets whose storytelling conventions differ from the CNN-style format we target.

5.2. Broader Impact

Automatically turning news video into structured articles has clear benefits. Qwen-NEWS can make video journalism more accessible to time-constrained readers, students, and researchers by exposing titles, highlights, and searchable text instead of requiring full playback. The enforced three-part structure preserves familiar editorial affordances: quick triage via headlines, bullet-pointed key facts, and paragraphs that provide context and chronology. For users with limited bandwidth or hearing impairments, text-based access can be more inclusive than streaming video.

At the same time, our system should complement rather than replace human journalists. The subtle multimodal hallucinations observed in Section 4.6 show that a high-scoring model can still invent scene details that are not supported by the transcript or keyframes. If such outputs were published without provenance or review, they could blur the boundary between documented events and model imagination. Automation also risks homogenizing narrative style as models converge on a small number of templated story patterns.

We therefore view Qwen-NEWS as a *decision-support* tool. Because the pipeline surfaces intermediate artifacts (Whisper transcripts and BLIP captions), editors can rapidly sanity-check generated articles against the underlying evidence, editing or rejecting sentences that lack support. This design emphasizes source traceability and helps maintain accountability in any human-in-the-loop deployment.

5.3. Future Directions

Several avenues could address these limitations. A first step is to move from simple concatenation to explicitly aligned multimodal reasoning. Instead of feeding transcripts and captions as two long blocks of text, future models could link transcript spans and keyframe captions along a shared timeline and condition generation on these alignments. Such designs would allow the LLMs to attend jointly over "who is speaking when" and "what is on screen at that moment," reducing the need to infer synchrony implicitly.

Second, visual grounding can be made an explicit objective rather than an emergent property. If sentences in the reference articles were annotated with supporting captions or timestamps, one could introduce an attribution-style loss that rewards claims backed by evidence and penalizes unsupported additions. In parallel, detecting low-variance visual streams at preprocessing time would allow the prompt to down-weight visual cues in studio-style clips, mitigating hallucinated crowds or cutaways of visually static video.

Third, evaluation should move beyond purely text-based similarity. Human studies that rate factuality, coherence, and perceived trustworthiness, together with claim-level checks against the transcript and keyframes, would more directly measure whether Qwen-NEWS is suitable for editorial workflows. On the automatic side, adapting recent evidence-based metrics to the video-to-text setting could provide finer-grained feedback on grounding.

Finally, deployment considerations motivate work on scaling and generalization. Distilling the LoRA-adapted 32B model into a smaller student would lower the compute barrier for newsrooms, while extending the dataset to multilingual and non-Western sources—leveraging Whisper's multilingual ASR—would test whether the same structured pipeline can support diverse journalistic traditions.

Together, these directions point toward multimodal summarization systems that provide fast, structured access to rich video content while maintaining stronger guarantees of factuality and evidence alignment.

6. Conclusion

We introduced Qwen-NEWS, a multimodal video-to-article framework that integrates Whisper-based ASR, BLIP keyframe captioning, and a LoRA-adapted Qwen-2.5-32B-Instruct backbone to generate structured news-style reports with Title, Highlights, and Article sections. On a carefully curated subset of MM-AVS, our system operates on dynamically reconstructed audio and visual streams and consistently outperforms strong prompt-only baselines across BERTScore, ROUGE, and METEOR. A training diagnostic further shows that LoRA fine-tuning yields a substantial reduction in negative log-likelihood on held-out data, indicating that even parameter-efficient adaptation is sufficient to specialize a general-purpose instruction-tuned model to the requirements of multimodal news summarization.

Qualitative analysis reveals that, when both the transcript and keyframe captions provide rich and complementary evidence, Qwen-NEWS produces fact-dense, temporally coherent articles that closely track human-written stories and adhere to the prescribed structure. At the same time, inspection of visually homogeneous clips exposes a characteristic failure mode in which the model hallucinates plausible but unsupported scene details, highlighting a gap between high text-based similarity scores and strict multimodal grounding.

In general, these findings suggest that structured video-to-text generation is well within the reach of current large language models, but that robust deployment will require tighter modeling of audio–visual alignment and evaluation protocols that explicitly account for evidence attribution and hallucination.

References

Chung, H. W., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1–53. https://arxiv.org/abs/2210.11416

Fu, X., Wang, J., & Yang, Z. (2021). MM-AVS: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5922–5926. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.473/

Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping Language-Image Pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*. https://arxiv.org/abs/2201.12086

Meena, P., Kumar, H., & Kumar Yadav, S. (2023). A review on video summarization techniques. *Engineering Applications of Artificial Intelligence*, 118, 105667. https://doi.org/10.1016/j.engappai.2022.105667

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. https://arxiv.org/abs/2212.04356

Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. https://jmlr.org/papers/volume21/20-074/20-074.pdf