Keyframe Selection and Multimodal Fusion for Product Recognition in E-Commerce Live Streaming

Yichuan Zheng , Jin Shi , Wei Shen *

Posted Date: 6 May 2026

doi: 10.20944/preprints202605.0313.v1

Keywords: e-commerce live streaming; product recognition; keyframe selection; multimodal fusion; large language models; object detection; image quality assessment; information extraction

 doi:10.20944/preprints202605.0313.v1

Article

Keyframe Selection and Multimodal Fusion for Product Recognition in E-Commerce Live Streaming

Yichuan Zheng [1], Jin Shi [2] and Wei Shen [1,*]

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310000, China
[2] China Jiliang University, Hangzhou 310000, China
* Correspondence: shenwei@zstu.edu.cn

Abstract

Product recognition in e-commerce live streaming is hindered by rapid viewpoint changes, occlusions, motion blur, and inconsistencies between visual and spoken information. Existing approaches typically focus on individual components such as detection, OCR, or speech recognition, which limits their effectiveness in end-to-end scenarios.To address this problem, we propose an integrated framework that combines task-oriented keyframe selection with multimodal semantic fusion. The framework first uses D-FINE to localize product regions, and then selects informative frames through two complementary strategies. Strategy A considers both detection confidence and Laplacian-based sharpness, while Strategy B combines detection confidence with a learned image-quality score estimated by an EfficientNetV2-based model. OCR, visual recognition, and ASR are then applied to the selected data, and a Qwen-Plus large language model is used to integrate multimodal evidence into structured product outputs. Experiments on an in-house dataset demonstrate significant gains over a last-frame baseline. Strategy A increases Perfect Match Rate from 58.00% to 80.00% and Product Name Recognition Accuracy from 78.00% to 98.00%. Strategy B achieves 77.00% and 98.00%, respectively. Ablation studies further show that the full multimodal framework consistently outperforms unimodal and dual-modality variants. In addition, Top-K analysis indicates that single-frame inference provides a good balance between performance and efficiency.Overall, the proposed framework offers an effective and practical solution for product recognition in complex live-streaming scenarios.

Keywords: e-commerce live streaming; product recognition; keyframe selection; multimodal fusion; large language models; object detection; image quality assessment; information extraction

1. Introduction

In recent years, e-commerce live streaming has emerged as a major driver of growth in the global retail industry, owing to its strong interactivity and high conversion rates. According to the 2024 China Live E-commerce Market Data Report released by Wangjing Society [1], the market size of China's e-commerce live-streaming sector exceeded RMB 5 trillion in 2024, with a user base surpassing 600 million. Internationally, live commerce has exhibited a similarly rapid growth trajectory. A report by Transparency Market Research indicates that the global live e-commerce market reached a value of USD 940.3 billion in 2024 and is projected to reach USD 6,079.8 billion by the end of 2035 [2]. This growth is particularly pronounced across major platforms and regions. For example, TikTok Shop achieved a global GMV of USD 32.6 billion in 2024, including USD 9.0 billion from the U.S. market [3]. In France, TikTok Shop's launch in March 2025 triggered rapid growth within six months: transaction volumes surged sevenfold, while livestream- and short-video-driven sales increased by 3.5 times and 14 times, respectively [4]. This expansion is supported by 27.8 million monthly active users in France and a broader European base exceeding 200 million. Similarly, Southeast Asia demonstrated robust growth in 2024, with Shopee and TikTok Shop achieving GMVs of USD 83.4 billion and USD 22.6 billion, respectively [5]. These figures indicate a steadily maturing global live-commerce ecosystem characterized by high interaction and strong conversion performance across major markets.

 doi:10.20944/preprints202605.0313.v1

The rapid expansion of this ecosystem has simultaneously increased the demand for scalable product understanding in live-streaming videos. As the primary carrier of product presentation and transaction-related evidence, livestream content requires systems that can reliably identify products at the semantic level rather than merely detect generic objects [6]. Such capability is important not only for merchandising and retrieval, but also for downstream applications including content review, compliance support, and consumer protection. In practice, however, the gap between coarse object localization and structured product recognition remains substantial, especially in unconstrained live-streaming environments.

The practical importance of this problem is amplified by the prevalence of counterfeit goods, misleading promotions, and visually ambiguous product presentations in live commerce. Recent enforcement reports and platform disclosures show that large-scale product screening remains a persistent challenge across major marketplaces [7–10]. This motivates recognition systems that are not only accurate at the image level, but also robust enough to support high-throughput video analysis in realistic operational settings.

Despite its importance, product recognition in live-streaming environments remains highly challenging. Frequent scene transitions, illumination changes, and occlusions make single-frame static methods unstable [11]. At the same time, streamers' spoken commentary and background noise introduce semantic redundancy and interference, which limits the reliability of single-modality approaches, whether based on vision or speech alone [6]. In addition, exhaustive analysis of the entire video stream incurs substantial computational cost, making practical deployment at platform scale difficult.

In response to these challenges, existing studies still exhibit an important gap. Much of the prior literature improves individual components, such as generic object detection, OCR, ASR, or pre-trained vision-language modeling [12], but comparatively little attention has been paid to end-to-end frameworks that are explicitly optimized for product recognition in e-commerce livestreams. In particular, there is still limited understanding of how to select recognition-effective frames from dynamic videos and how to reconcile complementary yet sometimes conflicting evidence across OCR, visual semantics, and speech.

To address this gap, we propose a keyframe extraction and multimodal fusion framework for structured product recognition in e-commerce live streaming. The central idea is to rank frames according to their utility for downstream recognition rather than according to generic visual quality alone. The framework therefore prioritizes frames that preserve task-effective information, such as packaging text, brand identifiers, and discriminative product appearance, and then integrates OCR, visual-semantic cues, and speech evidence through LLM-based semantic fusion. This design enables the system to produce standardized outputs for key product fields, including product name, brand, and category.

The main contributions of this work are threefold. First, we formulate keyframe selection for livestream product recognition as a task-oriented ranking problem and compare two practically motivated scoring strategies based on local sharpness and learned image-quality regression, respectively. Second, we develop an end-to-end multimodal framework in which OCR, ASR, and vision-language recognition are unified through structured evidence representation and LLM-centered semantic adjudication. Third, we provide a systematic empirical study on an in-house benchmark, including end-to-end comparisons, multimodal ablations, threshold sensitivity analysis, and single-frame versus Top-K trade-off evaluation. Collectively, these results show that effective frame selection and multimodal evidence fusion are both critical for robust product recognition in live-streaming scenarios.

2. Related Work

2.1. Overview of Related Work

To better support efficient and robust product information extraction and compliance verification for e-commerce live-streaming product recognition in regulatory scenarios, this section categorizes and

 doi:10.20944/preprints202605.0313.v1

reviews relevant existing studies based on the requirements of the proposed framework. Specifically, starting from the pain points of regulatory tasks, we systematically examine prior work across key components, including visual foundational detection (as the starting point for product localization), video dynamics processing and keyframe selection, image quality assessment (to ensure the reliability of visual modality inputs), multimodal understanding (for semantic evidence integration), and large language model–based fusion (to enable final adjudication). This organizational perspective not only traces the evolution and application of these techniques in e-commerce live streaming, but also highlights how they serve regulatory objectives through a purpose-driven progression—from low-level detection to high-level semantic fusion—thereby providing the theoretical foundation and motivation for the end-to-end solution proposed in this paper.

In the following subsections, Section 2.2 focuses on representative paradigms of visual object detection models (e.g., YOLO, DETR, and D-FINE), discussing their accuracy and real-time performance in product localization to motivate the detector selection in this work. Section 2.3 reviews multimodal approaches for video-level product recognition, analyzing their advances in advertising understanding while emphasizing their limitations in regulatory fine-grained tasks (e.g., product traceability), thereby motivating the need for multi-source evidence in our framework. Section 2.4 examines keyframe selection and multi-frame strategies, comparing how traditional information-theoretic and deep learning–based methods optimize computational efficiency, and revealing their disconnect from product semantic details (e.g., brand logos), which motivates the task-oriented scoring mechanism proposed in this study. Section 2.5 outlines the development of image quality assessment models, ranging from natural scene statistics to Transformer-based methods, and evaluates their role in enhancing the robustness of downstream recognition, while also pointing out the limitations of generic IQA in capturing task-relevant semantics, thereby motivating our joint modeling of image quality assessment and detection confidence. Finally, Section 2.6 explores the application of large language models to multimodal fusion, highlighting their potential for semantic reasoning while analyzing challenges related to output stability and engineering deployment, which supports the deterministic decoding and structured output strategy adopted in this work.

2.2. Review of Visual Object Detection Models

To justify the technical origin and selection rationale for the detection module, this subsection briefly reviews three representative detection paradigms and their typical implementations: the YOLO (You Only Look Once) family [13], DETR (Transformer-based end-to-end detection) [14], and the recently proposed D-FINE (a DETR-based fine-grained regression improvement) [15]. We then summarize their respective strengths and weaknesses and discuss their suitability for the task of product recognition in e-commerce video.

YOLO pioneered treating object detection as a single-shot regression problem, simultaneously predicting bounding boxes and class labels in one network forward pass, which enables high-frame-rate real-time detection [13]. This paradigm has been continuously optimized in subsequent engineering implementations such as YOLOv4/YOLOv5 (including data augmentation, network architecture, and training techniques), making the YOLO family a preferred baseline for industrial deployment and real-time applications (including retail shelf detection and video-stream detection). YOLO's advantages lie in high throughput and deployment friendliness, but one-stage methods typically lag behind more fine-grained approaches in localization accuracy for dense small-object scenarios or settings that demand very high bounding-box precision.

DETR reformulates object detection as a set prediction problem, employing a Transformer encoder–decoder [16] and the Hungarian matching loss [17] to produce end-to-end, non-redundant detection outputs, thereby obviating many hand-designed post-processing steps in traditional detectors (e.g., NMS [18], anchor design [19]) [14]. DETR's strengths include strong global-context modeling and a conceptually simple architecture; however, the original DETR suffers from slow convergence and somewhat weaker performance on small objects, issues that subsequent work has addressed with various improvements to accelerate training and strengthen local modeling. For video scenarios that

 doi:10.20944/preprints202605.0313.v1

require enhanced semantic understanding and global reasoning, DETR-style frameworks offer greater representational capacity.

D-FINE, recently proposed, reconstructs the boundary regression task within the DETR paradigm. Its two core innovations are Fine-grained Distribution Refinement (FDR) and Global Optimal Localization Self-Distillation (GO-LSD): FDR represents boundary regression as discrete probability distributions over distances to the four sides and iteratively refines them, thereby expressing localization uncertainty in a distributional form; GO-LSD leverages fine-level outputs to self-distill shallower layers, improving localization accuracy and training stability. [15] Experimental results show that D-FINE significantly improves localization precision while retaining the advantages of DETR, achieving a favorable trade-off between efficiency and accuracy and making it a strong candidate for tasks requiring precise localization (for example, recognizing product edges or packaging regions).

Based on the above comparison, the primary reasons for selecting D-FINE as the backbone detection module in this study are as follows: D-FINE augments the DETR framework with FDR and GO-LSD to substantially improve bounding-box localization precision and training stability while preserving the Transformer's global context modeling capability. In typical e-commerce live-streaming scenarios, products are often densely arranged, small in scale, subject to varying viewpoints, and affected by severe occlusion and motion blur. For example, in streamer demonstrations of cosmetics, snacks, or apparel, multiple small packaging boxes, lipsticks, or earrings may appear in the frame simultaneously. Compared with the YOLO family (where one-stage methods tend to underperform on dense small objects and localization precision) and the original DETR (which shows clear deficiencies in small-object detection and convergence speed), D-FINE has demonstrated superior AP on small-to-medium objects on benchmarks such as COCO, and performs especially well in fine-grained localization (e.g., product edges and packaging text regions) and complex occlusion scenarios.

Therefore, D-FINE's stable detection confidence and high-precision bounding-box outputs can provide more reliable product-region inputs for subsequent keyframe scoring and multimodal recognition, significantly improving overall recognition accuracy and robustness in dynamic, dense small-object e-commerce live-streaming scenarios and meeting regulatory requirements for precise capture of product details.

2.3. Video-Level Product Recognition and Multimodal Understanding

Early research on product recognition was predominantly based on detection and classification methods applied to static images. However, in video scenarios, factors such as dynamic content variations, shot transitions, occlusions, and motion blur make it difficult to directly transfer static-image-based approaches. To address these challenges, a growing body of recent work has incorporated both visual and speech modalities to enhance video advertisement understanding. For instance, in the context of structured analysis of advertisement videos, Weng et al. [20] proposed a multimodal framework that jointly leverages frame-level visual features and textual information extracted via OCR and ASR for scene segmentation and multimodal annotation, thereby improving the structured understanding of advertisement videos. This approach was also validated as an effective solution in the TAAC advertisement understanding challenge.

Although the aforementioned studies have explored joint modeling strategies of visual and speech modalities for video advertisement understanding, their primary focus lies in scene segmentation and content annotation, with limited attention paid to fine-grained product-level semantic recognition, which is more directly related to compliance supervision. Moreover, existing methods typically rely on fixed sampling strategies or full-video processing, without explicitly addressing frame quality variation and information sparsity that are prevalent in e-commerce scenarios. In contrast, the present work places greater emphasis on adaptively selecting information-rich key frames from dynamic videos and achieving high-accuracy product name recognition through multimodal fusion, thereby providing technical support for violation detection and authenticity verification in e-commerce livestreaming environments.

 doi:10.20944/preprints202605.0313.v1

2.4. Keyframe Selection and Multi-Frame Strategies

In the preprocessing stage of video recognition, key-frame selection and frame quality assessment play a crucial role in reducing computational cost and improving the performance of downstream recognition tasks. Traditional key-frame extraction methods often integrate multiple low-level visual features, such as color histograms, inter-frame differences, or motion estimation, to enhance selection quality. For example, Gu Jiayu et al. (2010) proposed a key-frame extraction method that combines MPEG-7 color layout descriptors with block-based motion information [21]. This approach captures the spatial distribution of colors using color layout features, analyzes local variations through block motion estimation, and adaptively extracts key frames via weighted fusion and cumulative distance measures. Other studies have adopted information-theoretic models to quantify frame importance. Cernekova et al. [22] utilized information entropy and mutual information to measure content variation between frames, enabling shot boundary detection and key-frame extraction by identifying information redundancy, thereby providing effective solutions for video summarization tasks.

However, the aforementioned methods primarily focus on global content changes and are often misaligned with the semantic requirements of downstream tasks such as product recognition. In the presence of occlusion, blur, or camera motion, the selected frames may not contain the most discriminative visual information (e.g., textual elements on product packaging or brand logos), which limits their applicability to fine-grained, task-oriented video understanding.

In recent years, with the development of no-reference image quality assessment (NR-IQA) [23] and deep learning techniques, researchers have increasingly employed learning-based models to predict frame quality and perform quality-aware frame selection, aiming to obtain more robust representative frames. Kharchevnikova and Savchenko [24] further proposed an efficient no-reference frame quality prediction approach using lightweight convolutional neural networks with teacher–student distillation, and demonstrated that selecting high-quality frames has a significant impact on downstream recognition performance. The frame scoring module in this work, which combines detection confidence with quality scores, is inspired by such approaches.

Regarding multi-frame strategies, prior studies have shown that a single frame is often insufficient to capture all useful information in a video. Consequently, selecting multiple frames and fusing their outputs can improve semantic coverage and robustness. For instance, some video summarization and key-frame extraction methods model frames as graph nodes and select representative frames based on semantic similarity or clustering mechanisms [25,26]. Nevertheless, these methods are primarily designed for summarization or visual abstraction, where the objective is to maximize semantic diversity and coverage across different scenes, rather than to ensure the reliable capture of critical semantic details such as brand names or packaging text.

In summary, existing key-frame selection and multi-frame strategies designed for video summarization or general video understanding exhibit the following core limitations when applied to product recognition as a downstream task:

(1) The absence of a quantifiable frame information value assessment mechanism tailored to downstream tasks;

(2) The inability to guarantee the preservation of task-relevant semantic details, such as textual information and brand identifiers;

(3) A focus in multi-frame strategies on semantic diversity rather than targeted optimization of product recognition performance.

Given that e-commerce product recognition highly depends on the clarity of local details and the completeness of visual information, this work adopts a task-oriented key-frame scoring mechanism that integrates product detection confidence with local sharpness and depth-related quality scores, thereby prioritizing frames with high recognizability and semantic value. Meanwhile, the proposed multi-frame strategy serves as an auxiliary mechanism to enhance information coverage, while the primary objective remains improving downstream multimodal recognition performance through the selection of high-value key frames.

 doi:10.20944/preprints202605.0313.v1

2.5. Image Quality Assessment (IQA)

In the field of image quality assessment (IQA), a large number of no-reference image quality assessment (NR-IQA) models have been proposed to predict subjective image quality in the absence of pristine reference images. Early approaches, such as BRISQUE [23] and NIQE [27], rely on natural scene statistics (NSS) to construct handcrafted features, which perform well on traditional distortion types but struggle to capture high-level semantic characteristics in complex scenes. With the rapid development of deep learning, convolutional neural networks (CNNs) have gradually become the dominant paradigm for NR-IQA. For example, HyperIQA [28] achieves more accurate quality prediction through a content-adaptive mechanism, while RankIQA [29] adopts a ranking-based learning strategy that learns feature representations from image quality rankings, effectively alleviating the scarcity of annotated data and improving regression performance. Subsequently, LIQE [30] approaches NR-IQA from a multi-task learning perspective by jointly optimizing quality assessment, scene classification, and distortion type recognition, and leverages the alignment capabilities of vision–language pre-trained models (CLIP) [31] to enable automatic knowledge transfer and loss re-weighting, further enhancing prediction accuracy and cross-dataset generalization. More recently, Vision Transformer (ViT)-based models [32], such as TReS [33] and MANIQA [34], have significantly improved the modeling of texture and structural distortions through global and multi-dimensional attention mechanisms, pushing NR-IQA performance to new levels.

However, generic NR-IQA models primarily focus on evaluating the perceptual quality of images (e.g., sharpness, noise, and contrast), and do not explicitly reflect the semantic quality required by downstream tasks. In tasks such as text recognition or product recognition, semantically informative content is often concentrated in local regions (e.g., brand logos or packaging text). As a result, an image with high overall perceptual quality but poor visibility of critical regions may be less favorable for recognition than a slightly blurred image that preserves complete semantic information. Therefore, directly adopting generic NR-IQA models may lead to a mismatch between quality scores and task performance.

Motivated by these limitations, this work introduces NR-IQA into the key-frame selection pipeline as an exploratory validation, aiming to examine whether selecting perceptually high-quality frames can facilitate product recognition performance. To compensate for the lack of semantic region modeling in conventional NR-IQA, the proposed method further combines quality scores with object detection confidence, performing image quality assessment within detected product bounding boxes. This joint modeling strategy allows both image quality and the visibility of task-relevant information to be considered simultaneously.

2.6. Large Language Models and Multimodal Fusion

In recent years, the rapid advancement of large language models (LLMs) has provided more powerful semantic understanding and reasoning capabilities for multimodal fusion. Existing studies primarily focus on architectural innovations, such as designing global–local attention mechanisms to integrate OCR, ASR, and visual features (e.g., the multimodal Transformer proposed by Tsai et al. [12]), or encoding visual information into natural language descriptions that are subsequently processed by LLMs for open-ended reasoning (e.g., the BLIP-2 framework proposed by Li et al. [35]). While these approaches enhance the depth of multimodal understanding, they often suffer from unstable outputs and a lack of structured representations in practical applications, which limits their direct applicability to e-commerce video recognition scenarios that demand high reliability and reproducibility.

The limitations of existing methods are mainly reflected in the following aspects: high-temperature sampling or open-ended generation strategies may lead to semantic drift or variations in expression for identical inputs, making it difficult to meet the stringent requirements for reproducibility and auditability in regulatory settings. Moreover, these methods typically produce free-form natural language outputs without standardized structured fields (e.g., product name, brand, or category), which necessitates additional parsing or manual intervention and substantially increases system

 doi:10.20944/preprints202605.0313.v1

deployment costs. In large-scale e-commerce livestreaming regulation, platforms are required to conduct real-time compliance review over massive volumes of video content, where any output instability or missing fields may result in missed detections or false judgments, thereby undermining the efficiency of screening for false advertising or counterfeit products.

Motivated by these engineering challenges, this work proposes a fusion paradigm that emphasizes practicality and stability. Built upon the Qwen series models, the proposed approach employs carefully designed prompt templates and deterministic decoding strategies (with temperature T=0) to achieve stable and structured multimodal outputs on top of strong foundation models. The goal is to provide an efficient, reliable, and readily deployable semantic fusion solution for e-commerce video recognition systems.

3. Method

3.1. Problem Definition

The objective of this study is to automatically extract and recognize product information from e-commerce livestreaming videos to support platform-level content regulation, such as false advertising detection and counterfeit product screening. Given a livestreaming video $V = \{f_1, f_2, \ldots, f_n\}$, where f_i denotes the i-th video frame, along with the corresponding audio stream A, the task is defined as follows.

Input: the video stream V and audio stream A, together with a predefined set of target product fields $\mathcal{F} = \{\text{product name}, \text{brand}, \text{product category}\}$.

Output: a structured product information set $\mathcal{G} = \{(g_1, g_2, \ldots, g_m)\}$, where each $g_j \in \mathcal{F}$ represents the predicted value of the corresponding field along with its confidence score.

Constraints: (1) resource-aware and practical deployment requirements, where processing latency should remain within an acceptable range for long live-streaming videos; (2) high reliability, where recognition results must be traceable and verifiable.

The core challenge arises from the highly dynamic nature of livestreaming videos, in which product appearances are transient and highly variable. This necessitates the intelligent selection of an information-rich frame subset $K \subseteq \{1, 2, \ldots, n\}$ from V, followed by the integration of visual (OCR and Qwen-VL) and auditory (ASR) modality evidence. A large language model is then employed to perform semantic-level decision making and output accurate, structured product information $\mathcal{G}$. The objective of this work is to minimize the end-to-end recognition error while ensuring that computational efficiency satisfies platform-scale deployment requirements, formulated as:

$$\min_{K,\Theta} \mathcal{L}(\mathcal{G}\text{pred}(K, \Theta), \mathcal{G}\text{gt}) \quad \text{s.t.} \quad \text{Time}(V) \leq T_{\max} \tag{1}$$

where Θ denotes the system parameters and $\mathcal{L}$ represents the end-to-end loss function.

This study focuses on optimizing the visual modality in e-commerce livestreaming scenarios. Specifically, we investigate how to automatically select frames from the video stream that provide the maximum amount of informative content for downstream recognition, namely frames that are most valuable in terms of visual clarity, the presence of key information (e.g., packaging text and brand logos), and detection confidence. The goal is to enable the OCR and image recognition stages (e.g., Qwen-VL) to obtain as much accurate evidence as possible. Rather than improving the performance of individual modules in isolation, the ultimate objective is to achieve sufficient and accurate product recognition at the end-to-end level through key-frame selection, unified structured outputs of multimodal evidence, and semantic-level fusion based on large language models. To this end, this work designs and compares frame scoring strategies based on traditional visual features and deep quality regression, as well as single-frame and Top-K multi-frame fusion strategies, with end-to-end recognition performance on real-world e-commerce video datasets serving as the final evaluation criterion.

 doi:10.20944/preprints202605.0313.v1

3.2. Overall Architecture

The overall workflow of the proposed system is illustrated in Figure 1. The proposed system takes videos as input and selects representative frames (hereafter referred to as candidate frames or selected frames) through a frame extraction and frame quality scoring module. For each selected frame, three modality-specific recognition pipelines are executed in parallel: OCR (PaddleOCR [36]), ASR (FunASR [37]), and product image recognition (Qwen-VL). The raw recognition outputs from each pipeline are subsequently structured and standardized by a large language model (LLM), producing unified-format evidence units. All evidence units are then aggregated and fed into a final fusion LLM, which performs multimodal reasoning and outputs the final recognition results along with supporting evidence.

- **Frame extraction & scoring**: Frames are sampled from the input video at a fixed frame rate, and a quality score is computed for each frame (e.g., a combination of detection confidence and Laplacian sharpness). According to the scoring strategy, a single frame or multiple candidate frames are selected for downstream processing.
- **OCR pipeline**: Text recognition is performed on the candidate frames to obtain raw textual content and internal OCR confidence scores. The raw text, confidence scores, and positional information are provided as input to an LLM, which is instructed to return normalized and structured outputs, such as brand name, product name, product category, and specifications.
- **ASR pipeline**: The system performs speech recognition on the entire audio stream of the video to obtain a complete transcription of spoken content. This transcription does not include temporal or frame-level alignment information; instead of localizing specific video segments, it serves as semantic evidence that assists in understanding the products presented in the video. The LLM then parses and standardizes the raw ASR transcription to produce structured information, including potentially mentioned product names, brands, specifications, and their semantic confidence scores. At this stage, the LLM effectively conducts a semantic-level preliminary judgment based on spoken descriptions, forming textual evidence units parallel to the visual modality.
- **Visual recognition with Qwen-VL**: For each product region detected in the candidate frames, Qwen-VL is invoked to perform image-based recognition and description. The model input consists of cropped image regions and task-specific prompts, and the output includes product category, brand, specifications, and confidence scores, forming standardized structured evidence units.
- **Evidence aggregation and fusion LLM**: Structured outputs from the three pipelines are aggregated as contextual input and provided to the fusion LLM with explicit fusion instructions. The fusion LLM produces the final decision, including the product name, brand, and product category.

 doi:10.20944/preprints202605.0313.v1

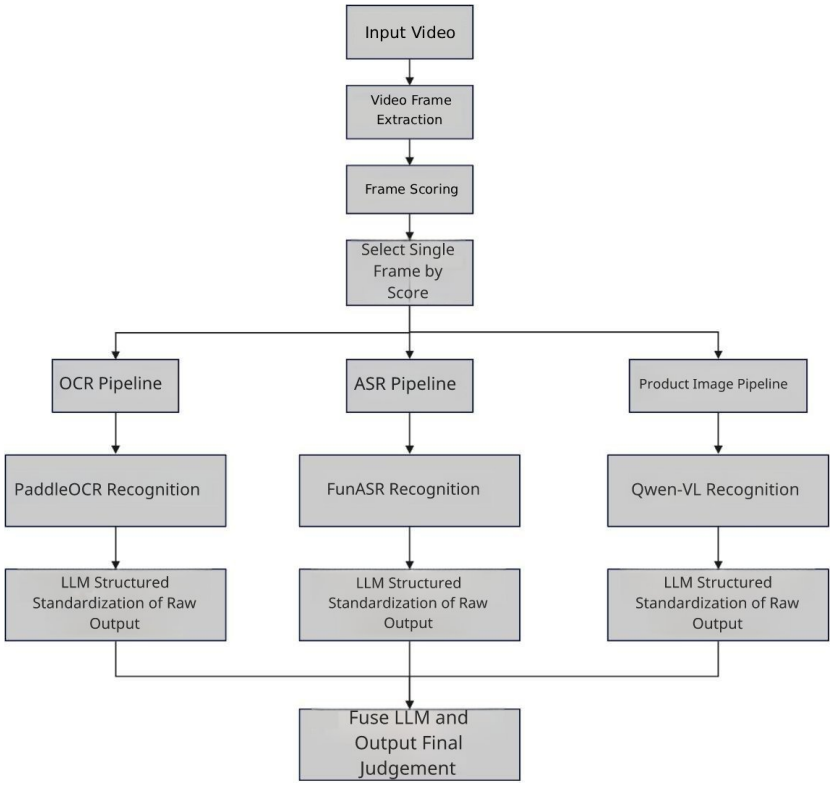


Figure 1. Overall architecture of the proposed model.

3.3. Frame Scoring Strategy

To select the most suitable frames from video streams for product recognition, this study designs and compares three types of frame scoring strategies: a Baseline method (fixed-position sampling), Strategy A (based on traditional visual features), and Strategy B (based on deep learning–based quality regression). The selected frames are subsequently used as inputs for OCR and multimodal recognition, thereby reducing the computational cost of downstream recognition modules while improving overall recognition accuracy. The following sections describe the definitions, implementation details, and boundary-case handling mechanisms of the three strategies in turn.

3.3.1. Baseline: Fixed-Position Strategy

The Baseline strategy simply selects the last frame of the video as the recognition frame without performing any quality assessment, serving as a reference baseline. This method has the lowest computational complexity; however, in live-streaming scenarios, the final frame is often a transition or scene-switching frame, which may significantly degrade recognition performance. Therefore, it is used primarily as a benchmark for evaluating the effectiveness of downstream methods.

3.3.2. Strategy A: Traditional Visual Feature–Based Scoring

Strategy A assigns a quality score to each frame by jointly considering object detection confidence and image sharpness. For each frame f_i in the video, the overall quality score is defined as:

$$S_i = w_d \cdot C_d(f_i) + w_s \cdot C_s(f_i), \tag{2}$$

where $C_d(f_i) \in [0,1]$ denotes the detection confidence score, $C_s(f_i) \in [0,1]$ denotes the sharpness score, and the weights w_d and w_s satisfy $w_d + w_s = 1$. Unless otherwise specified, this work adopts $w_d = 0.7$ and $w_s = 0.3$.

Detection confidence (C_d): When multiple detection boxes are present in a frame, $C_d(f_i)$ is computed as the arithmetic mean of the confidence scores of all detected product bounding boxes. Alternatively, top-k averaging or maximum-value strategies may be employed to emphasize the most

 doi:10.20944/preprints202605.0313.v1

reliable candidate. In this work, mean aggregation is used to mitigate the impact of occasional false detections with abnormally high confidence. Since the detector outputs confidence scores already normalized to the range $[0, 1]$, no additional normalization is required.

Sharpness score (C_s): For each detected product region of interest (ROI), image sharpness is evaluated using the response of the Laplacian operator and normalized as:

$$C_s(f_i) = \min\left(\frac{\mathrm{Var}(\Delta^2 ROI_i)}{500}, 1.0\right), \tag{3}$$

where Δ^2 denotes the Laplacian operator, $\mathrm{Var}(\cdot)$ represents the variance operation, ROI_i corresponds to a detected product region, and 500 is an empirically chosen normalization factor.

When multiple product regions are present in a frame, the frame-level sharpness score $C_s(f_i)$ is obtained by averaging the sharpness values of all ROI_i. If the detector fails to return any valid ROI, the Laplacian variance computed over the entire frame is used as a degraded fallback estimate of $C_s(f_i)$ to avoid missing values.

Overall, this strategy is computationally efficient and highly interpretable, making it suitable for practical frame scoring in resource-constrained deployment scenarios.

3.3.3. Strategy B: Deep Learning–Based Scoring

Strategy B replaces the traditional operator-based sharpness metric with a trained deep regression model Q_θ to estimate the *recognizability* score of an image. The frame scoring function is defined as:

$$S_i = w_d \cdot C_d(f_i) + w_q \cdot Q_\theta(f_i), \tag{4}$$

where $Q_\theta(f_i) \in [0, 1]$ denotes the regression output of an EfficientNetV2-M model, and the weights w_d and w_q satisfy $w_d + w_q = 1$. Unless otherwise specified, this work adopts $w_d = 0.7$ and $w_q = 0.3$.

Model output normalization. After training, the raw regression outputs of Q_θ are linearly normalized to the range $[0, 1]$ to ensure numerical comparability with the sharpness scores used in Strategy A.

Input preprocessing. For each detected product region of interest (ROI), the corresponding image patch is cropped and resized to 224×224 pixels before being fed into the regression model. If no detection bounding boxes are present in a frame, the entire frame is used as the model input to avoid missing scores.

Training procedure. The regression model is trained using pseudo-labels generated by a teacher model (HyperIQA). Supervised learning is performed by minimizing the mean squared error (MSE) loss between the predicted scores and the pseudo-labels. After training, the model is further evaluated on a validation set by examining the correlation between its predicted scores and the actual OCR recognition success rate, in order to verify its effectiveness for downstream recognition tasks.

Overall, Strategy B generally outperforms the operator-based Strategy A under complex noise and motion-blur conditions, at the cost of additional model inference overhead.

3.4. Frame Selection and Implementation Details

3.4.1. Frame Selection Pipeline

Given an input video $V = \{f_1, \ldots, f_n\}$, the frame selection procedure is performed as follows:

- A detector is applied to each frame to obtain product bounding boxes and their confidence scores.
- For each detected bounding box, either a sharpness score is computed (Strategy A) or the cropped region is fed into the regression model Q_θ (Strategy B).
- A frame-level composite score S_i is calculated for each frame.
- The frame with the highest composite score is selected as the final recognition frame:

$$f_{\text{best}} = \arg\max_i S_i. \tag{5}$$

 doi:10.20944/preprints202605.0313.v1

- The selected frame is then passed to the downstream OCR and multimodal fusion modules.

3.4.2. Top-K Multi-Frame Fusion Strategy

To further improve robustness and information completeness, this work also explores an optional Top-K multi-frame fusion strategy. Instead of selecting a single frame, the top K frames ($K > 1$) with the highest quality scores are selected from the video, OCR is performed on each frame independently, and the recognition results are fused at the text level.

Specifically, all frames are first sorted in descending order according to their quality scores. A greedy selection algorithm with a minimum inter-frame distance constraint is then applied: frames are examined sequentially, and a candidate frame is skipped if its temporal distance to any already selected frame is smaller than a predefined threshold. This constraint prevents selecting temporally adjacent and visually redundant frames. By leveraging complementary information across frames—such as variations in viewpoint, illumination, and occlusion—this strategy can effectively improve both recall and recognition accuracy.

The fusion of multi-frame OCR results consists of four steps. First, all OCR texts are normalized, including lowercasing, removal of redundant punctuation, unification of numeric formats and units, and application of common OCR error correction rules, in order to reduce superficial textual discrepancies. Second, deduplication and fuzzy matching are performed: exact matching is prioritized, while approximately similar but non-identical strings are merged based on string similarity metrics. Third, for texts merged into the same candidate item, the maximum confidence score among all sources is retained as the final confidence. Finally, unmerged independent text items are preserved to capture complementary information across frames, yielding a more comprehensive and reliable recognition result.

3.5. Multimodal Fusion and Final Matching

To achieve robust product recognition in videos, this work constructs a large language model (LLM)-based multimodal fusion framework. The framework integrates complementary information from visual modalities (OCR and image recognition) and the auditory modality (ASR), represents them uniformly in textual form, and feeds them into an LLM, which performs semantic-level fusion and reasoning to generate the final product recognition results.

3.5.1. Multimodal Information Extraction

For the optimal frame selected by the frame scoring strategy, the system simultaneously extracts information from three modalities:

Visual–Text Modality (OCR). PaddleOCR is applied to the selected product frame to recognize textual information appearing on the product packaging. The outputs include recognized text content, spatial coordinates, and confidence scores. This modality primarily captures salient product identifiers such as product names, brands, specifications, ingredients, and promotional slogans, and thus serves as a direct source of explicit identification information.

Auditory Modality (ASR). FunASR is used to perform speech recognition on the entire audio stream of the video, producing a transcription of spoken content. The ASR modality extracts product-related information from semantic descriptions, such as the host's introduction of product features, usage scenarios, promotional language, and price descriptions. Through these semantic cues, product names, brands, and specifications can be inferred, effectively complementing limitations of purely visual information.

Visual–Semantic Modality (Image Recognition). The Qwen-VL multimodal model is employed to perform image recognition and description on detected product regions. The model outputs predicted product categories, candidate labels, and natural language descriptions. This modality provides appearance-based semantic features—such as packaging style and functional type—and can offer supplementary judgments when textual information is missing or poorly recognized.

 doi:10.20944/preprints202605.0313.v1

The recognition results from the three modalities are respectively converted into textual representations denoted as T_{OCR}, T_{ASR}, and T_{Image}, and stored in a structured format for subsequent multimodal fusion.

3.5.2. LLM-Based Fusion Mechanism

In this work, the Qwen-Plus large language model is adopted as the core module for multimodal fusion. The model input is constructed by concatenating the textual outputs from the three modalities, while a system prompt explicitly defines the task objective, fusion logic, and output format.

The fusion process follows the principles below:

1. **Consistency First**: When multiple modalities provide identical or highly consistent information, such results are prioritized to enhance the reliability of the final output.
2. **Reliability Ranking**: Modalities are fused according to their confidence-based reliability, with the priority order defined as OCR (packaging text) > image recognition > ASR (speech), ensuring that highly reliable evidence dominates the final decision.
3. **Conflict Resolution**: In cases where outputs from different modalities conflict, the result with higher confidence and stronger consistency with OCR evidence is preferred.
4. **Fault-Tolerant Completion**: When information from certain modalities is missing, the model generates a complete judgment based on the available modalities, thereby maintaining the stability and robustness of the system output.

The final fusion result generated by the LLM is structured into the following standardized format:

$$Q_{\text{fusion}} = \{\text{Product Name, Brand, Category, Key Attributes, Price}\}$$

3.5.3. Temperature Parameter and Inference Stability

During the inference stage of large language models, the temperature parameter T is a key hyperparameter that controls the randomness of generation and has a direct impact on output stability and reproducibility. In this work, T is consistently set to 0 across all experiments, corresponding to a greedy decoding strategy, where the model selects only the token with the highest probability at each step, thereby minimizing randomness in the generation process. The probability distribution can be formulated as:

$$P(w_t \mid w_{<t}) = \text{softmax}\left(\frac{z_t}{T}\right), \tag{6}$$

where z_t denotes the predicted logits at time step t. When the temperature is high, the output distribution becomes smoother, and the model samples among multiple candidate tokens, increasing generation diversity. In contrast, as the temperature approaches zero, the distribution becomes increasingly peaked, and the model deterministically selects the most probable token at each step.

In this study, fixing the temperature parameter to zero is motivated by three main considerations. First, a low temperature ensures the reliability and comparability of experimental results. When comparing different frame scoring strategies or multimodal fusion configurations, stochastic generation may lead to divergent outputs for identical inputs, introducing additional variance that can obscure experimental conclusions. By adopting deterministic decoding, the model produces identical outputs under the same input conditions, significantly improving reproducibility and verifiability. Second, setting $T = 0$ helps maintain the stability of structured output formats. Since the LLM outputs in this work are constrained to predefined fields (e.g., product name, brand, category, key attributes, and price), higher temperatures may result in deviations from the template or missing fields, complicating automated evaluation and downstream matching. Greedy decoding ensures consistent structure and complete fields, facilitating subsequent programmatic analysis. Finally, a zero temperature reduces the risk of semantic drift. Multimodal fusion requires the model to preserve semantic consistency and logical coherence when integrating evidence from multiple sources, whereas higher temperatures may cause the model to deviate from the input semantics and generate redundant or irrelevant content.

 doi:10.20944/preprints202605.0313.v1

Fixing the temperature to zero reinforces faithful semantic alignment with the input evidence, thereby improving the interpretability and reliability of the generated results.

3.6. Data

3.6.1. Data Sources and Overall Statistics

This study constructs a dedicated dataset for product recognition in e-commerce livestream scenarios. The dataset is constructed from publicly available e-commerce livestream videos sourced from major platforms, including Taobao, Douyin, and Pinduoduo. To capture a wide spectrum of live-streaming environments, such as varying streamer styles, camera perspectives, and UI layouts, a total of 442 training videos were collected.From these videos, approximately 36,000 raw frames were extracted at a fixed frame rate. After manual filtering, annotation, and data augmentation, around 3,000 images were obtained for training the product detection model (D-FINE). The raw frames were further processed by the trained product detector, resulting in approximately 11,000 frames containing visible products.

To support end-to-end evaluation, an additional evaluation set was constructed, consisting of 100 manually annotated product entries corresponding to 100 video samples, which are used for validating the final recognition performance.

3.6.2. Annotation Protocol and Workflow

Annotation targets. To improve the generalization ability of the detector with respect to product "recognizability," the D-FINE model adopts a single unified category, *goods*, for annotation. Only bounding boxes of product instances are labeled, without distinguishing specific product categories or brands.

Annotation criteria. Annotators are required to outline the complete and clearly recognizable external contour of each product. Objects that are severely occluded, extremely low-resolution, or visually unrecognizable are excluded from annotation.

Annotation records. Each annotation file includes fields such as image_id, bbox, annotator_id, and annotator_confidence, enabling subsequent traceability and annotation quality analysis.

3.6.3. Data Augmentation and Preprocessing

To enhance the generalization performance of D-FINE under limited data conditions, a set of linear-combination-based data augmentation strategies was applied to the manually annotated samples. These include horizontal and vertical flipping, brightness/contrast/saturation perturbations, random cropping, and mild Gaussian noise injection. The augmentation process expands the original annotated dataset by approximately 2×, resulting in around 3,000 samples used for D-FINE training. All input images are normalized and standardized prior to training.

3.6.4. Pseudo-Labeling Pipeline for EfficientNet Training

The EfficientNet model is trained to predict image "recognizability" or quality scores. Due to the high cost of manual quality annotation, a semi-supervised pseudo-labeling strategy is adopted. Specifically, approximately 36,000 frames are first extracted from the 442 videos, from which about 11,000 product-containing frames are retained using the product detector. These frames are then scored using HyperIQA, a no-reference image quality assessment model, and the resulting scores are used as pseudo-labels for training the EfficientNet-based regression model.

3.6.5. Subset Sampling for Comparative Experiments

Frame scoring comparison. For comparing frame scoring strategies A and B, 50 videos are randomly sampled from the 442-video dataset using a fixed random seed (seed = 42). Both scoring strategies are applied to these videos, and the selected frames are evaluated in terms of OCR recognizability, average OCR confidence, number of recognized text instances, and processing latency.

 doi:10.20944/preprints202605.0313.v1

End-to-end recognition evaluation. End-to-end experiments are conducted on the set of 100 manually annotated product videos. System outputs are matched against manually standardized product names using BGE-M3 embedding similarity for recognition correctness determination. All sampling procedures and random operations are documented in the paper, including random seeds, sampling counts, and implementation scripts, to ensure reproducibility.

3.7. Models and Training

This study trains two core models: (1) **D-FINE**, which is used for detecting products in video frames and outputs product bounding boxes along with detection confidence scores; and (2) **EfficientNet**, which is employed to assess image "recognizability/quality" and provides the quality metric for the frame scoring strategy (Strategy B). Both models are trained on a self-constructed dataset, and all data splits are performed at the video level to ensure that the training, validation, and test sets do not overlap.

3.7.1. D-FINE Detector

For candidate product region detection, we adopt an implementation of the D-FINE framework (Detection with Fine-grained Localization and Global Self-Distillation) to improve bounding box localization accuracy and detection robustness in livestream video scenarios. The model consists of three main components: an HGNetv2-B2 backbone, a HybridEncoder, and a DFINETransformer decoder. The backbone is initialized with official pretrained weights to obtain strong general-purpose visual representations. The encoder aligns multi-scale features into a unified channel dimension (hidden_dim = 256) via multi-scale feature fusion. The decoder follows a Transformer architecture with deformable attention, comprising four decoding layers (num_layers = 4) and 300 query vectors (num_queries = 300). It directly outputs detection boxes and class predictions in an end-to-end manner, without relying on conventional FPN or NMS post-processing.

The detector predicts two classes (product/background) and is used solely for detection; it does not output embeddings for semantic retrieval. Subsequent semantic recognition is performed by the OCR and multimodal recognition modules.

The core innovations of D-FINE lie in Fine-grained Distribution Refinement (FDR) and Global Optimal Localization Self-Distillation (GO-LSD) [15]. FDR replaces direct numerical regression of bounding box coordinates with discrete distribution modeling, where the model predicts a probability distribution for each boundary position and progressively refines it in a residual manner across decoding layers, enabling fine-grained modeling of spatial uncertainty. GO-LSD treats the localization results from the final decoding layer as teacher signals and distills them into intermediate layers using a distribution consistency loss, thereby enforcing globally optimal inter-layer constraints. This mechanism is implemented through the Fine-grained Localization (FGL) loss and the Decoupled Distillation Focal (DDF) loss, where FGL enforces unimodal distribution constraints and boundary refinement, while DDF focuses on inter-layer distillation.

By jointly considering classification confidence, bounding box regression, distribution prediction, and distillation constraints, the overall training objective is defined as:

$$L = 1.0 \times L_{\text{vfl}} + 5.0 \times L_{\text{bbox}} + 2.0 \times L_{\text{giou}} + 0.15 \times L_{\text{fgl}} + 1.5 \times L_{\text{ddf}} \tag{7}$$

Here, L_{vfl} denotes the Varifocal Loss for confidence-weighted classification optimization, L_{bbox} is the L1 bounding box regression loss, L_{giou} is the GIoU loss, and L_{fgl} and L_{ddf} correspond to the fine-grained localization and distillation constraint terms, respectively. This weighting strategy has been shown in the original D-FINE work to significantly improve localization accuracy for small and medium-sized objects as well as convergence speed.

Training configuration and hyperparameters: The input resolution is fixed at 640×640 pixels, and training is conducted using COCO-format annotations. The AdamW optimizer is employed, together with a linear warm-up phase and a multi-stage learning rate decay schedule to balance

 doi:10.20944/preprints202605.0313.v1

convergence speed and training stability. Mixed-precision training (AMP) and Exponential Moving Average (EMA) are enabled to improve generalization performance and numerical stability. Data augmentation strategies include brightness and contrast perturbations, random scaling, IoU-based cropping, and horizontal flipping, enhancing robustness to illumination changes and viewpoint variations. The main hyperparameters and training strategies are summarized in Table 1.

Table 1. Training Hyperparameters of the D-FINE Detector.

Category	Parameter	Value
Model	Backbone	HGNetv2-B2 (pretrained)
	Hidden dimension	256
	Decoder layers	4
	Number of queries	300
	Number of classes	2 (product, background)
Training	Epochs	132
	Batch size	16
	Learning rate	2.5×10^{-4} (backbone: 2.5×10^{-5})
	Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1.25×10^{-4})
	LR schedule	MultiStepLR (milestone=500, $\gamma = 0.1$) + linear warmup
	Warmup steps	500
	AMP	Enabled
	EMA	Enabled (decay=0.9999)
Data	Input size	640×640
	Data augmentation	PhotometricDistort, RandomZoomOut, RandomIoUCrop, Flip
	Stop epoch	120
Loss	Loss weights	vfl: 1.0, bbox: 5.0, giou: 2.0, fgl: 0.15, ddf: 1.5
Evaluation	Metrics	COCO mAP (IoU 0.5–0.95), AP50, AP75

Optimization Objective and Evaluation: The model's evaluation metrics follow the COCO-style standards, including mAP@[0.5:0.95], AP50, and AP75. Training logs and validation results are recorded at each epoch, and early stopping is applied to select the best model when no improvement is observed in the validation set metrics. The final output confidence scores are used for subsequent frame scoring and candidate prioritization.

Advantages: Compared to traditional two-stage detection frameworks (e.g., Faster R-CNN + FPN), the DETR-style end-to-end detection in D-FINE offers three main advantages:

Elimination of the NMS post-processing step, achieving true end-to-end optimization.

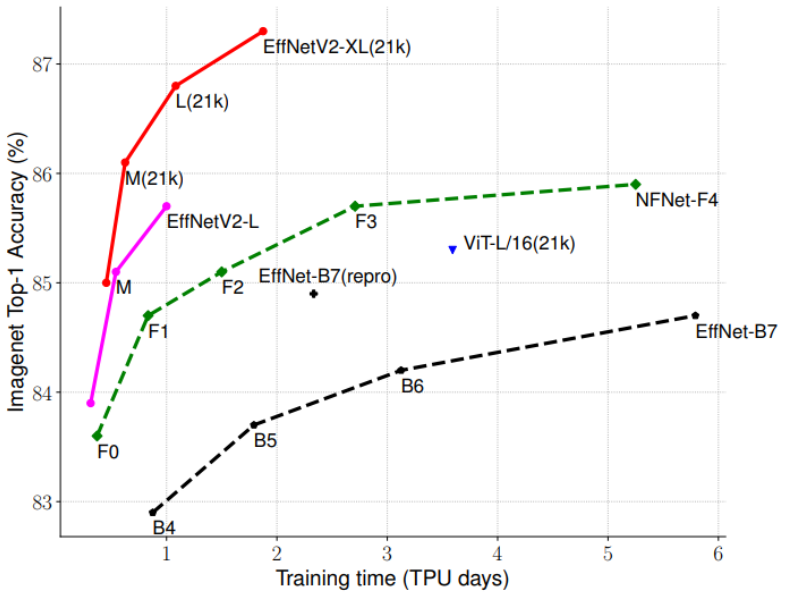
The FDR + GO-LSD mechanism significantly improves bounding box accuracy, showing stronger robustness, especially in live-streaming product scenes with large scale changes and motion blur.

The Transformer decoder structure is more compatible with downstream multimodal models (e.g., Qwen-VL, LLM), providing structural compatibility for unified vision-language reasoning.

3.7.2. EfficientNetV2

In the frame scoring (Strategy B), we use EfficientNetV2-M as the "image recognizability regressor" to estimate the recognizability score of candidate frames. This score is then normalized and used as the quality term in the frame ranking. The choice of EfficientNetV2-M as the backbone is based on its advantages in training efficiency and parameter efficiency: compared to traditional ConvNets or some Transformer variants, EfficientNetV2 significantly reduces training time and inference latency while maintaining high accuracy, as illustrated in the performance comparison in Figure 2 [38].

 doi:10.20944/preprints202605.0313.v1



(a) Training efficiency.

	EfficientNet (2019)	ResNet-RS (2021)	DeiT/ViT (2021)	EfficientNetV2 (ours)
Top-1 Acc.	84.3%	84.0%	83.1%	83.9%
Parameters	43M	164M	86M	24M

(b) Parameter efficiency.

Figure 2. Training speed/parameter scale comparison of EfficientNetV2 (Source: Mingxing Tan et al., 2021 [38]).

This strategy is inspired by Kharchevnikova & Savchenko (2021) [24], who used lightweight CNNs combined with distillation/weak-labeling to build an efficient frame quality estimator for video frame selection/quality assessment. The approach generates "quality" pseudo-labels for a large number of unlabeled or weakly-labeled frames using a teacher model, and then a student model learns this quality scale for cost-effective large-scale training. Based on this idea, in the live product recognition scenario, we use a no-reference image quality evaluator (HyperIQA) as the teacher model to generate continuous quality scores for the candidate frames extracted at a fixed frame rate and filtered by a detector. These pseudo-labels are then used to train EfficientNetV2-M, allowing it to learn to predict the "recognizability" score correlated with downstream recognition success rates. The goal of this design is to approximate the teacher's scores using a lightweight student model with minimal overhead during inference, thus providing an efficient and deployable quality metric in frame scoring (Strategy B).

Let the output of EfficientNetV2-M be defined as a continuous regression value $\hat{y}$, with the training objective being to minimize the mean squared error (MSE) between the predicted value and the teacher's pseudo-label y:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{8}$$

The training data consists of approximately 36,000 frames extracted from 442 videos at a fixed frame rate, and after filtering by the detector, about 11,000 frames containing products. The data is split by video into training and validation sets (80% / 20%). During training, input images undergo uniform preprocessing and data augmentation (see Table 2). The ReduceLROnPlateau learning rate scheduler and early stopping mechanism are employed to avoid overfitting, and mixed precision training (FP16) is enabled to accelerate training and reduce memory usage.

 doi:10.20944/preprints202605.0313.v1

After training, the regression output is linearly normalized to the range $[0, 1]$ as the final quality score. The regression metrics (MSE, MAE, RMSE) are reported on the validation set, along with the Pearson correlation coefficient between the model's predictions and the teacher's scores.

Table 2. EfficientNet Training Hyperparameters.

Category	Parameter	Value
Model	Backbone	EfficientNetV2-M (pretrained on ImageNet-1K)
	Input resolution	224×224
	Parameters	$\approx$ 24M
	Dropout rate	0.3
Training	Batch size	16
	Epochs	20
	Learning rate	5×10^{-4}
	Optimizer	Adam
	Weight decay	1×10^{-4}
Loss	Function	MSE (regression)
LR schedule	Scheduler	ReduceLROnPlateau (factor=0.5, patience=3, min lr 1×10^{-5})
Data	Split	80% / 20%
Augmentation	Methods	Flip (50%), brightness/contrast ($\pm$30%), Gaussian noise (20%)
Training strategy	Early stopping	5 epochs

4. Experiments

This section aims to validate the effectiveness of the proposed frame scoring strategy, Top-K multi-frame fusion mechanism, and multimodal recognition framework. The experiments are conducted from three perspectives: frame selection quality evaluation, OCR information extraction performance comparison, and end-to-end recognition accuracy analysis. All experiments are performed under a unified system environment, ensuring consistency in model parameters, inference settings, and hardware conditions to guarantee comparability and reproducibility of the results.

4.1. Model Training Results and Analysis

4.1.1. D-FINE Training Results and Performance Analysis

The model training process converges smoothly, reaching its final performance after 132 epochs. The specific performance metrics are shown in Table 3.

Table 3. Performance Metrics of the D-FINE Detector.

Metric	Value	Description
mAP (IoU 0.5–0.95)	27.76%	COCO standard mean Average Precision over IoU thresholds from 0.5 to 0.95
AP_{50}	40.36%	Average Precision at IoU = 0.5
AP_{75}	28.25%	Average Precision at IoU = 0.75
AR_{100}	44.14%	Average Recall with up to 100 detections per image
Final training loss	20.71	Loss value at the end of training (epoch 131)

Given the relatively small scale of the training set and the task difficulty (product detection requires precise localization), the final mAP@0.5:0.95 of 27.76% and mAP@0.5 of 40.36% are at a reasonable level. Compared to models trained on large-scale datasets (e.g., COCO with over 80 categories and more than 120k training images), the performance of this model on a small dataset demonstrates:

effective transfer learning: leveraging Objects365 pre-trained weights enables rapid convergence on the small dataset;

successful domain adaptation:despite limited training samples, the model successfully learns product detection features;

 doi:10.20944/preprints202605.0313.v1

practical orientation: the 44.14% recall rate ensures that the vast majority of frames containing products can be detected in frame selection applications.

4.1.2. EfficientNetV2 Training Results and Performance Analysis

The model converges well on both the training and validation sets, triggering the early stopping mechanism after reaching the best validation performance at epoch 17. The final model performance is shown in Table 4:

Table 4. Performance Metrics of EfficientNetV2.

Metric	Value	Description
Validation MSE	0.00220	Mean Squared Error
Validation MAE	0.03667	Mean Absolute Error
Validation RMSE	0.04687	Root Mean Squared Error
Pearson correlation	0.9323	Linear correlation between predicted scores and ground-truth scores
Best epoch	17	Epoch with the lowest validation loss

The model's Pearson correlation coefficient reaches 0.9323, indicating that the predicted quality scores are highly consistent with the scores from the HyperIQA teacher model, successfully achieving knowledge distillation. The MAE of 0.0367 corresponds to an average error of only 3.67% within the [0,1] scoring range, demonstrating high prediction accuracy. The MSE and MAE are relatively close, with RMSE slightly higher than MAE, suggesting the presence of a small number of outlier samples with larger errors, but overall predictions remain stable.

4.2. Experimental Metrics and Statistical Methods

To comprehensively evaluate the performance across different experiments, this paper defines corresponding metric systems for each experimental setting.

4.2.1. Experiment 1: OCR Information Quantity Metrics

This experiment evaluates OCR performance from three aspects: recognition quantity, confidence, and efficiency, thereby assessing whether multi-frame fusion effectively improves the amount and reliability of recognized information. The specific metrics are shown in Table 5.

Table 5. Evaluation Metrics for Experiment 1

Metric	Description
Avg. number of characters	Average number of characters recognized per video
Avg. confidence	Mean confidence score of all recognized characters
High-confidence characters	Average number of characters per video with confidence ≥ 0.8
Processing time	Average processing time per video from input to OCR completion

4.2.2. Experiment 2: Frame Quality Evaluation Metrics

This experiment employs downstream task performance (OCR recognition results) as a unified objective evaluation benchmark. By assessing the performance of selected frames in terms of text recognition, detection confidence, and efficiency, it indirectly measures the superiority of the two strategies. The specific metrics are shown in Table 6.

 doi:10.20944/preprints202605.0313.v1

Table 6. Frame Quality Evaluation Metrics for Experiment 2

Metric	Definition	Role
OCR Average Confidence	Average confidence of all recognized characters in the selected frames	Reflects the clarity and readability of text regions
Detection Confidence	Average confidence of detection boxes from D-FINE	Measures the detectability of targets in the image and overall quality
Proportion of High-Confidence Text	Proportion of text with confidence no less than 0.7	Evaluates the proportion of high-quality recognition results
OCR Text Quantity	Total number of recognized text entries	Reflects the completeness of information extraction
Processing Time	Total time from video input to output of selected frames	Reflects the overall runtime efficiency of the algorithm

In addition, the experiment includes a human evaluation component to complement the objective metrics. Five annotators independently scored the representative frames selected by different strategies from four aspects: text clarity, product completeness, brand visibility, and overall visual quality. A five-point scale was adopted, where 1 indicates very poor quality, 2 poor quality, 3 acceptable quality, 4 good quality, and 5 excellent quality. More specifically, a score of 1 corresponds to frames with severe blur, incomplete product presentation, unclear brand cues, or almost unreadable text; a score of 3 corresponds to frames that are basically usable but still contain noticeable defects; and a score of 5 corresponds to frames with clear text, complete product presentation, clearly visible brand information, and high overall readability. To reduce subjective bias, the candidate frames were presented in randomized order without revealing the corresponding strategy labels. For each sample, the scores given by all annotators were first averaged within each evaluation dimension, and the final human evaluation score was then obtained by averaging the dimension-level mean scores. This subjective evaluation is combined with the quantitative metrics to verify the consistency between objective performance and human perception across different strategies.

4.2.3. Experiment 3: Recognition Accuracy Metrics and Threshold Setting

This experiment focuses on the end-to-end product recognition task, comparing the overall performance of different frame selection strategies (Baseline, Strategy A, and Strategy B) in the multimodal fusion recognition stage. To ensure a comprehensive evaluation, this paper assesses recognition performance at both the field level and the video level.

1. **Field-Level Accuracy**
   Semantic similarity between recognition results and ground-truth annotations is calculated for each field, with the following definitions:
   - **Brand Recognition Accuracy**: proportion of samples with brand similarity ≥ 0.7;
   - **Product Name Recognition Accuracy**: proportion of samples with product-name similarity ≥ 0.5;
   - **Category Recognition Accuracy**: proportion of samples with category similarity ≥ 0.5.
2. **Video-Level Accuracy**
   Based on the matching results of the three fields, two video-level evaluation metrics are defined:
   - **Perfect Match Rate**: proportion of videos whose brand similarity ≥ 0.7, product-name similarity ≥ 0.5, and category similarity ≥ 0.5 simultaneously;
   - **Semantic Similarity**: average semantic similarity of the three fields, namely brand, product name, and category.

Semantic similarity is computed using text embeddings generated by the pre-trained BGE-M3 model and measured by cosine similarity. Compared with exact string matching, this metric better reflects whether the predicted result is semantically consistent with the ground-truth annotation.

 doi:10.20944/preprints202605.0313.v1

The default thresholds are determined according to the semantic characteristics of different fields. Specifically, the brand field usually contains short and semantically explicit expressions, so a stricter threshold of 0.7 is adopted. In contrast, product names in e-commerce live-streaming scenarios often include abbreviations, specification supplements, and promotional modifiers, which introduce substantial expression variations; therefore, a relatively tolerant threshold of 0.5 is used. The category field is comparatively coarse-grained, and a threshold of 0.5 is also adopted to preserve moderate semantic tolerance.

To validate the rationality of the adopted thresholds, a one-factor sensitivity analysis is conducted by varying one threshold at a time while fixing the other two. Since threshold variation only affects the correctness criterion rather than the semantic similarity itself, Table 7 presents the changes in Perfect Match Rate and the corresponding field accuracy under different threshold settings.

Table 7. Sensitivity Analysis of Evaluation Thresholds

Variable	Threshold	Perfect Match Rate (%)	Corresponding Field Accuracy (%)
Brand	0.60	81.00	81.00
Brand	0.65	80.00	80.00
Brand	**0.70**	**80.00**	**80.00**
Brand	0.75	78.00	78.00
Brand	0.80	74.00	74.00
Product Name	0.40	80.00	100.00
Product Name	0.45	80.00	99.00
Product Name	**0.50**	**80.00**	**98.00**
Product Name	0.55	79.00	97.00
Product Name	0.60	75.00	90.00
Product Name	0.65	72.00	84.00
Product Name	0.70	67.00	76.00
Category	0.40	80.00	99.00
Category	0.45	80.00	99.00
Category	**0.50**	**80.00**	**99.00**
Category	0.55	74.00	88.00
Category	0.60	66.00	77.00

As shown in Table 7, the selected thresholds yield comparatively stable results across different settings. For the brand field, the performance varies only slightly around the default threshold, indicating that 0.7 is an appropriate choice for a semantically explicit field. For the product-name field, increasing the threshold above 0.5 results in a clear decrease in both Perfect Match Rate and the corresponding field accuracy, suggesting that 0.5 is better suited to product names with frequent expression variations. For the category field, the results remain stable when the threshold is no higher than 0.5, whereas higher thresholds lead to a noticeable decline in performance. These results support the final threshold setting of 0.7 for brand, 0.5 for product name, and 0.5 for category.

4.3. Experiment 1: Comparison Between Single-Frame and Top-K Multi-Frame Fusion

4.3.1. Experimental Setup

This experiment aims to validate the improvement effect of the Top-*K* multi-frame fusion strategy on text recognition (OCR) results. The compared strategies include the following three:

1. **Single-Frame Strategy**: Directly select the single frame with the highest comprehensive quality score as the OCR input;
2. **Top-3 Fusion Strategy**: Select the top 3 frames with the highest quality scores and fuse their OCR outputs;

 doi:10.20944/preprints202605.0313.v1

3. **Top-5 Fusion Strategy**: Select the top 5 frames with the highest quality scores and fuse their OCR outputs.

30 videos are randomly sampled for the comparison experiment. All strategies employ the same frame scoring method, namely Strategy A, computed through a weighted combination of detection confidence (weight 0.7) and Laplacian clarity (weight 0.3). To avoid redundant information from temporally adjacent frames, a minimum inter-frame distance constraint is introduced during frame selection:

$$d_{\min} = \left\lfloor \frac{n}{2K} \right\rfloor,$$

ensuring uniform distribution of selected frames along the timeline.

In the multi-frame fusion stage, a "text deduplication + confidence weighting" approach is used to integrate OCR outputs from each frame, thereby obtaining the final text recognition results.

4.3.2. Experimental Results

The performance of single-frame and multi-frame strategies is compared according to the metrics for Experiment 1 defined in Section 4.2. The experimental results are shown in Table 8.

Table 8. Comparison of OCR Performance Under Different Numbers of Selected Frames.

Metric	Single Frame ($K = 1$)	Top-3 Frames	Top-5 Frames
Average Number of Characters	17	20	25
Average Confidence	0.7593	0.8897	0.8850
Average Number of High-Confidence Characters	15	18	20
Average Processing Time (s)	32.51	93.11	155.52

4.3.3. Results Analysis

As shown in Table 8, multi-frame fusion outperforms the single-frame strategy in OCR performance. The Top-3 fusion achieves an approximately 17.2% improvement in average confidence compared to the single frame and a 20% increase in the number of high-confidence characters, indicating that multi-frame information can indeed compensate for blur or missing content in single frames to a certain extent. However, this performance gain comes with significant time overhead: the processing time for the Top-3 strategy is approximately 2.9 times that of the single-frame strategy, while Top-5 reaches 4.8 times.

Furthermore, the performance improvement from Top-3 to Top-5 is very limited, with only an 11% increase in high-confidence characters and even a slight decrease in average confidence (from 0.8897 to 0.8850), suggesting that beyond a certain number of frames, redundant information begins to diminish the fusion effectiveness. Considering the trade-off between accuracy improvement and efficiency cost, the overall benefit of multi-frame fusion is limited, particularly making it less suitable for practical deployment in regulatory scenarios with strict efficiency constraints.

Therefore, this study ultimately selects the single-frame strategy as the default for the system. This strategy ensures relatively high recognition accuracy while significantly reducing computational cost and response latency, better aligning with lightweight and resource-aware deployment requirements. The multi-frame fusion strategy can be retained as an optional extension in the future for specific scenarios with extremely high demands on recognition accuracy.

4.4. Experiment 2: Comparison of Frame Quality Between Strategy A and Strategy B

4.4.1. Experimental Setup

This experiment aims to compare the differences in frame selection performance between Strategy A, based on traditional computer vision methods, and Strategy B, based on deep learning quality assessment. Since the internal comprehensive scoring mechanisms of the two strategies differ—Strategy A uses linear weighting of D-FINE detection confidence and Laplacian clarity, while Strategy B uses detection confidence and quality scores output by the EfficientNetV2-M model—their score scales and

 doi:10.20944/preprints202605.0313.v1

distributions are not directly comparable. Therefore, this experiment does not directly compare scores but instead adopts downstream task performance (OCR recognition results) as a unified evaluation criterion. By analyzing the performance of selected frames in the text recognition stage, it indirectly reflects the differences between the frame selection strategies in terms of image quality and readability.

The overall procedure of this experiment is as follows:

1. **Input Data Preparation**: Randomly sample 50 videos from 442 product-related videos as the test set to ensure diversity in scenes and content.
2. **Frame Selection Stage**: Apply Strategy A and Strategy B separately to score and rank frames for each video, selecting the highest-scoring single frame as the representative frame for each strategy. The weight parameters of both strategies are kept consistent, with detection confidence weighted at 0.7 and quality score weighted at 0.3, to ensure fairness in the comparison.
3. **OCR Recognition Stage**: Use the same OCR model (PaddleOCR) to perform text recognition on the selected representative frames, extracting recognized text and corresponding confidence scores. This experiment evaluates only the frame performance in the OCR stage, without incorporating ASR or image recognition modules. This is because the core objective of the frame scoring strategy is to select clear frames with sufficient information, and its effectiveness should primarily be reflected through the recognition performance of visible text.
4. **Results Statistics and Analysis**: Perform statistical analysis on the OCR recognition results of both strategies across all test videos, including metrics such as average confidence, recognized text quantity, proportion of high-confidence text, and frame index differences.

4.4.2. Experimental Results

This experiment conducts a comprehensive comparison and evaluation of Strategy A and Strategy B, combining automatic and human evaluation.

Automatic Evaluation Results

(1) OCR Confidence Performance

As shown in Figure 3, the average OCR confidence for Strategy A is 0.7665 ± 0.1618, while for Strategy B it is 0.8037 ± 0.1113. Compared to Strategy A, Strategy B improves the average confidence by $+0.0373$ (relative improvement of 4.86%). Among the 50 test samples, Strategy B achieves higher confidence than Strategy A in 48.0% (24/50) of the samples.

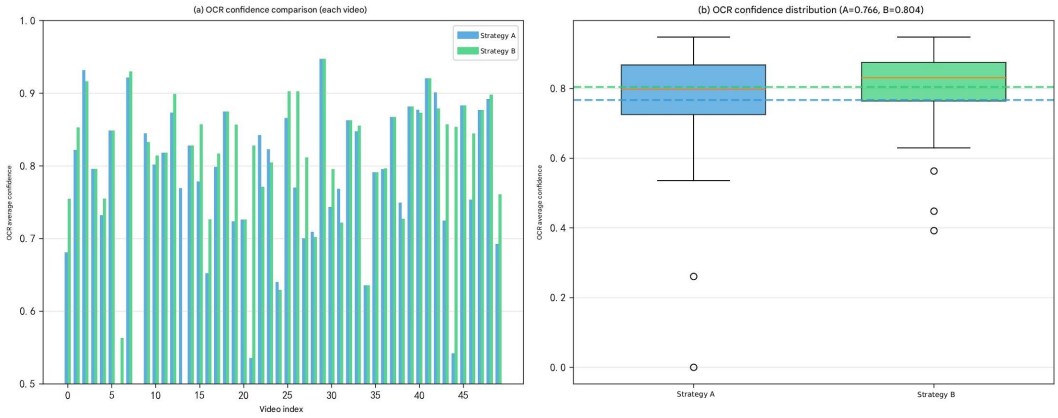


Figure 3. Comparison of OCR Confidence

Notably, Strategy B not only increases the average confidence but also significantly reduces the standard deviation (from 0.1618 to 0.1113), with the minimum confidence rising from 0 to 0.3920, indicating superior performance of Strategy B in avoiding extremely low-quality frames.

(2) Text Recognition Quality

 doi:10.20944/preprints202605.0313.v1

As shown in Figures 4 and 5, the proportion of high-confidence text for Strategy A is 72.22%, while for Strategy B it is 74.88%; the average total text length for Strategy A is 131.3 characters, and for Strategy B it is 130.2 characters.

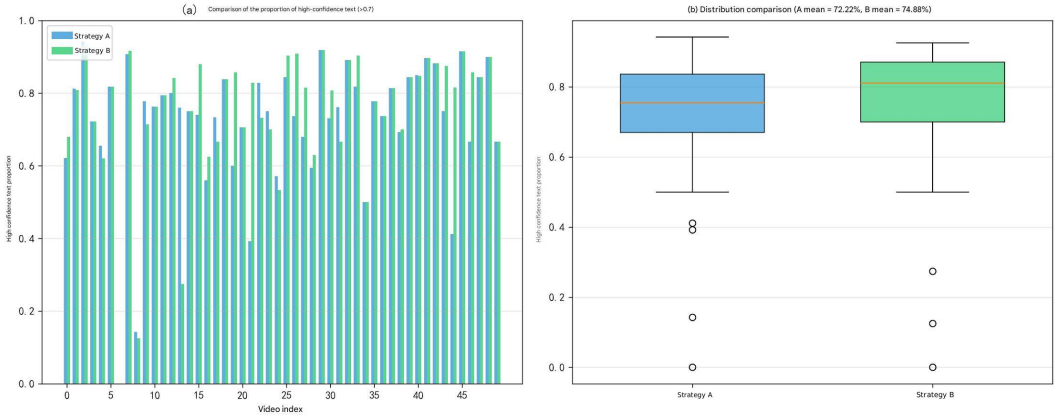


Figure 4. Comparison of Proportion of High-Confidence Text

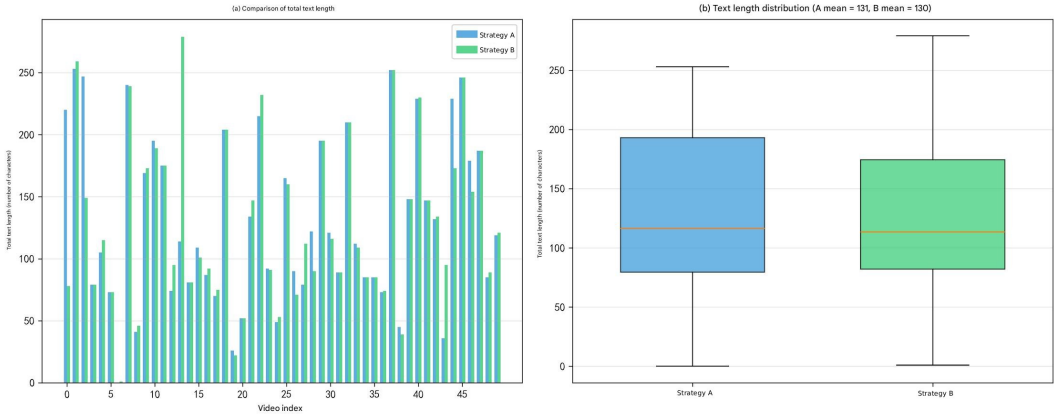


Figure 5. Comparison of Average Text Length

Although the text quantities are similar, Strategy B successfully increases the proportion of high-confidence text, which is crucial for subsequent information extraction.

(3) Improvement and Degradation Performance

Among the 50 samples, the distribution of performance changes for Strategy B relative to Strategy A is shown in Table 9.

Table 9. Performance Change Distribution Under Different Strategies.

	Improvement	Degradation	Equivalent
Number of Samples	24 (48.0%)	11 (22.0%)	15 (30.0%)

(4) Time Cost

As shown in Figure 6, the average processing time for Strategy A is 105.14 seconds, while for Strategy B it is 113.43 seconds, an increase of 8.29 seconds (+7.9%).

 doi:10.20944/preprints202605.0313.v1

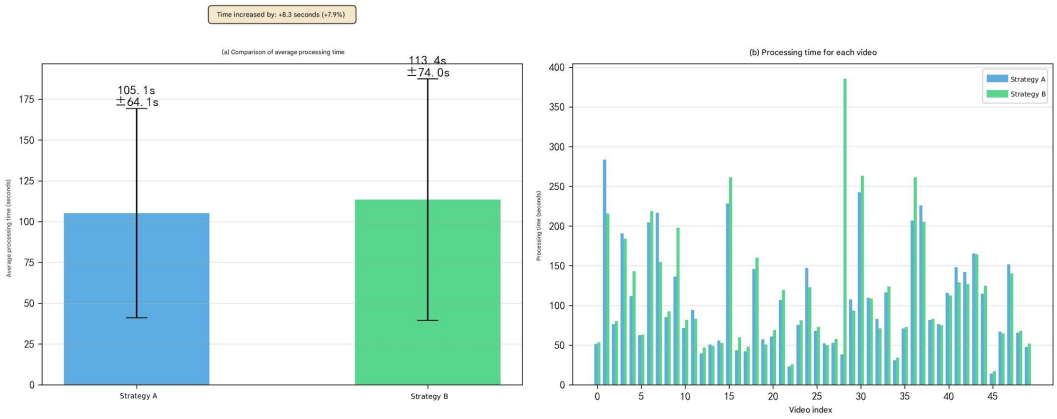


Figure 6. Comparison of Processing Time

Human Evaluation Results

To complement the automatic evaluation, this study further conducts human evaluation on the 50 samples, and the results are shown in Table 10.

Table 10. Comparison of Human Evaluation Results for Strategy A and Strategy B.

	Strategy A	Strategy B	Score Improvement (B relative to A)
Average Quality Score	3.43	3.73	+0.30 (8.75%)

In the human evaluation, Strategy B receives higher scores than Strategy A in 38.0% (19/50) of the samples, while Strategy A receives higher scores in 10.0% (5/50). The remaining 52.0% (26/50) of the samples obtain comparable scores. This result is consistent with the automatic evaluation trend, further supporting the advantage of Strategy B in subjective frame quality assessment.

4.4.3. Results Analysis

The experimental results demonstrate that Strategy B achieves the expected improvements in frame quality assessment:

(1) Improved OCR Confidence and Stability

By incorporating an image quality assessment model, Strategy B considers not only OCR text quantity but also comprehensive image quality factors such as clarity and contrast when selecting candidate frames. Experimental data show that Strategy B improves average OCR confidence by 4.86%, and more importantly, reduces the standard deviation of confidence by 31.2% (from 0.1618 to 0.1113), with the minimum confidence rising from 0 to 0.3920. This indicates that image quality assessment effectively filters out low-quality frames with blur, severe occlusion, or poor illumination.

(2) Avoidance of Extremely Low-Quality Frames

In detailed case analysis, for product 06901443453603, the frame selected by Strategy A has an OCR confidence of 0 (complete failure to recognize), while the frame selected by Strategy B achieves a confidence of 0.563, representing a qualitative leap from "complete failure" to "basically usable." Similar improvement cases account for 48% of the total samples, proving the value of image quality assessment in risk mitigation and effectiveness.

(3) Human Evaluation Validation

The human evaluation results are generally consistent with the automatic evaluation. Strategy B improves the average quality score by 8.75% and is rated higher than Strategy A in 38.0% of the samples, while Strategy A is rated higher in 10.0% of the samples. The remaining 52.0% of the samples receive comparable scores, indicating that both strategies can select frames of acceptable quality in a considerable proportion of cases.

 doi:10.20944/preprints202605.0313.v1

4.5. Experiment 3: End-to-End Recognition Accuracy Comparison

4.5.1. Experimental Setup

This experiment serves as the end-to-end validation of this paper, aiming to evaluate the comprehensive impact of different frame selection strategies (Baseline, Strategy A, Strategy B) on the final product recognition accuracy within the complete multimodal recognition pipeline, thereby verifying the effectiveness of the proposed multimodal fusion framework in real-world product recognition tasks. The compared strategies are as follows:

1. **Baseline**: Directly use the last frame of the video for recognition;
2. **Strategy A**: Perform end-to-end recognition using the frame selection Strategy A proposed in Experiment 2;
3. **Strategy B**: Perform end-to-end recognition using the frame selection Strategy B proposed in Experiment 2.

The complete experimental pipeline includes: frame selection $\rightarrow$ running OCR (PaddleOCR) and image recognition (Qwen-VL) on the selected frame $\rightarrow$ running ASR (FunASR) on the entire video $\rightarrow$ feeding the structured evidence units from the three modalities into a large language model (Qwen-Plus, temperature $T = 0$) for fusion reasoning $\rightarrow$ using BGE-M3 to perform field-level semantic similarity matching and judgment between LLM outputs and ground-truth annotations.

By comparing the final semantic similarity results of different frame scoring strategies in the multimodal recognition pipeline, the impact on product recognition accuracy and information completeness is assessed.

4.5.2. Experimental Results

In the end-to-end evaluation on 100 samples, the recognition accuracies of three strategies—Baseline (directly using the last frame of the video), Strategy A (based on traditional CV methods), and Strategy B (incorporating an image quality assessment model for frame selection)—are compared. The quantitative results are shown in Table 11:

Table 11. Performance Comparison of Different Frame Scoring Strategies.

Metric	Baseline	Strategy A	Strategy B
Perfect Match Rate	58.00%	80.00%	77.00%
Product Name Recognition Accuracy	78.00%	98.00%	98.00%
Semantic Similarity	68.75%	82.38%	80.40%

4.5.3. Results Analysis

(1) Overall Performance Analysis

As shown in the end-to-end results in Table 11, the proposed "key frame selection + multimodal fusion" framework achieves significant and practically meaningful improvements in real-world product recognition tasks. Taking the perfect match rate as an example, compared to the baseline (58.00%), Strategy A improves to 80.00% (absolute improvement of 22.00%, relative improvement of 37.9%), and Strategy B to 77.00% (absolute improvement of 19.00%, relative improvement of 32.8%). This improvement is already substantial in practical product recognition scenarios, capable of significantly reducing manual verification and user correction costs.

At a finer granularity, the product name recognition accuracy improves from 78.00% in the baseline to 98.00% in both Strategy A and Strategy B (absolute improvement of 20.00%). Achieving 98% recognition accuracy for product names indicates that the system can provide precise product semantic information in the vast majority of videos. The semantic similarity also improves from 68.75% to 82.38% / 80.40% (Strategy A/B), reflecting that multimodal fusion significantly enhances output consistency and reliability at the semantic level. In summary, high-quality frame inputs (obtained through frame

 doi:10.20944/preprints202605.0313.v1

selection) and reliable information integration (through LLM fusion of OCR/ASR/Qwen-VL) jointly drive the end-to-end performance improvement.

(2) Analysis of Why Strategy B Underperforms Strategy A in the Full Pipeline

Although Strategy B performs better in frame quality and downstream OCR tasks (average OCR confidence 0.8037 vs. 0.7665, proportion of high-confidence text 74.88% vs. 72.22%), it still slightly underperforms Strategy A in end-to-end recognition accuracy. In the present 100-sample evaluation, although Strategy B reaches the same product name recognition accuracy as Strategy A (98.00%), its perfect match rate (77.00%) and semantic similarity (80.40%) remain lower than those of Strategy A (80.00% and 82.38%, respectively). Possible reasons for this phenomenon are as follows.

On one hand, there is the constraint of information completeness. In the end-to-end recognition pipeline, if key information (such as brand logo or complete product name) is invisible or incomplete in the selected frame, subsequent visual processing cannot proceed effectively regardless of OCR quality. Experimental data show that Strategy A recognizes an average of 131.3 characters, while Strategy B recognizes 130.2 characters, a difference of only 1.1 characters (0.8%). However, these 1.1 characters may precisely contain critical brand identifiers or specification information. More importantly, in pursuit of image quality, Strategy B may select clear but information-incomplete frames (e.g., clear side or back views), leading to missing frontal brand logos or complete product names.

On the other hand, there is incomplete adaptation of the teacher model in Strategy B. The image quality assessment model (HyperIQA) is trained on natural image quality datasets, with evaluation criteria including general visual quality metrics such as clarity, contrast, color saturation, and noise level. However, product recognition scenarios have particular characteristics: brand logos and product names are typically concentrated in specific positions on the packaging (front, top labels) rather than uniformly distributed; a frame with high global quality score may not have optimal quality and visibility in key information regions; products are three-dimensional objects, and different shooting angles display completely different information—a compositionally elegant, extremely clear side view may receive a high quality score, but its information value is far lower than a slightly blurred but front-facing photo; in some cases, frames with uniform illumination and moderate contrast receive high quality scores, but overly uniform lighting may cause loss of discriminability in reflective materials (e.g., metallic texture of brand logos), making recognition more difficult.

4.6. Experiment 4: Ablation Study on Multimodal Evidence Fusion

4.6.1. Experimental Setup

To quantify the contribution of different modalities, an ablation study is conducted on an end-to-end evaluation set comprising 100 video samples. For each sample, a Ground Truth database containing core fields (brand, product name, specifications) was manually constructed as the reference.

The experiment fixes the keyframe selection to Strategy A and compares four configurations:

- OCR only: Uses only textual information from keyframes.
- OCR+ASR: Adds speech transcripts to supplement visual text.
- OCR+Qwen-VL: Combines text with visual-semantic features from the VLM.
- OCR+ASR+Qwen-VL: The proposed full multimodal framework.

All groups use the same LLM adjudicator with deterministic decoding ($T = 0$) to ensure that performance variances stem solely from the modality combinations.

4.6.2. Experimental Results

Three metrics are reported in this experiment: Perfect Match Rate, Product Name Recognition Accuracy, and Semantic Similarity. The results are summarized in Table 12.

 doi:10.20944/preprints202605.0313.v1

Table 12. Performance Comparison of Modality Combinations.

Modality	Perfect Match Rate	Product Name Recognition Accuracy	Semantic Similarity
OCR only	70.00%	92.00%	78.41%
OCR+ASR	75.00%	97.00%	80.72%
OCR+Qwen-VL	76.00%	96.00%	81.09%
Full Framework	**80.00%**	**98.00%**	**82.38%**

4.6.3. Results Analysis

As shown in Table 12, using OCR alone already achieves relatively strong performance, with a Perfect Match Rate of 70.00%, Product Name Recognition Accuracy of 92.00%, and Semantic Similarity of 78.41%. This indicates that textual information on product packaging remains the most direct and reliable source in live-streaming product recognition. It also validates the effectiveness of the keyframe selection module, which prioritizes the readability of textual regions to provide high-quality inputs for subsequent recognition.

With the introduction of speech information, the OCR+ASR combination improves the Perfect Match Rate to 75.00%, Product Name Recognition Accuracy to 97.00%, and Semantic Similarity to 80.72%. This improvement mainly stems from the complementary role of speech signals. In live-streaming scenarios, product names and key attributes are often explicitly mentioned by streamers, allowing ASR to compensate for missing or incomplete visual text, especially under occlusion or blur.

In contrast, OCR+Qwen-VL, as an alternative dual-modality configuration, achieves higher performance in Perfect Match Rate (76.00%) and Semantic Similarity (81.09%) compared to OCR+ASR. This demonstrates that the visual-semantic information provided by Qwen-VL—such as brand logos, packaging appearance, and category-specific features—effectively enhances overall semantic understanding. However, its Product Name Recognition Accuracy (96.00%) is slightly lower than that of OCR+ASR, suggesting that visual semantics are less effective than speech signals for fine-grained product name recovery, but more advantageous in improving semantic consistency.

When all three modalities are combined, the OCR+ASR+Qwen-VL framework achieves the best overall performance, with a Perfect Match Rate of 80.00%, Product Name Recognition Accuracy of 98.00%, and Semantic Similarity of 82.38%. Compared to the OCR-only baseline, these correspond to improvements of 10.00%, 6.00%, and 3.97%, respectively. This result indicates a strong complementary relationship among the modalities: OCR provides explicit textual evidence, ASR contributes speech-based semantic information, and Qwen-VL offers visual-semantic cues. Their integration enables the system to construct a more complete cross-modal evidence chain, leading to improved recognition accuracy and semantic consistency.

In summary, the ablation results confirm the foundational role of OCR and demonstrate that ASR and Qwen-VL provide complementary enhancements from speech and visual-semantic perspectives, respectively. Their combination yields the best performance across all metrics, validating the effectiveness of the proposed multimodal fusion framework for product recognition in live-streaming scenarios.

5. Conclusions

This paper presents an end-to-end framework for structured product recognition in e-commerce live streaming. The framework combines task-oriented keyframe selection with multimodal semantic fusion, enabling the system to preserve recognition-effective visual evidence while integrating OCR, ASR, and visual-semantic cues at the decision level. Relative to conventional pipelines that either rely on fixed-position frame sampling or process video streams exhaustively, the proposed design provides a more favorable balance between recognition quality and computational efficiency.

In the system evaluation, we systematically compared the single-frame and Top-K multi-frame strategies, as well as two frame scoring strategies (Strategy A: detection confidence + local clarity; Strategy B: detection confidence + deep quality regression). The experimental results show that:

 doi:10.20944/preprints202605.0313.v1

- The Top-K multi-frame fusion strategy provides significant improvements in information completeness (including the number of recognized OCR characters and high-confidence text), but its computational cost increases substantially with larger K. Considering both efficiency and recognition accuracy, the single-frame strategy offers higher engineering feasibility in resource-constrained and practical deployment scenarios.
- Strategy B outperforms Strategy A in OCR confidence and frame selection stability, more effectively avoiding the selection of extremely low-quality frames. However, in the final end-to-end evaluation, Strategy A still achieves better overall recognition performance, with higher Perfect Match Rate and Semantic Similarity than Strategy B. This result reflects a trade-off between "global image quality" and "task-relevant information completeness." Based on the above findings, this paper recommends Strategy A as the default single-frame strategy for practical deployment, while Strategy B or Top-K multi-frame fusion can be used as supplements in scenarios with looser efficiency constraints or in applications with higher requirements for recognition confidence.

Compared with prior work, the main distinction of this study lies in treating frame selection as a recognition-oriented problem rather than a generic video summarization or image-quality assessment problem. The results further show that multimodal gains are not obtained merely by adding more inputs; instead, they depend on selecting frames that preserve task-effective information and on reconciling heterogeneous evidence through semantic-level fusion. From this perspective, the proposed framework offers a practically grounded solution for product recognition in dynamic livestream environments while also clarifying the trade-off between global image quality and downstream semantic completeness.

Despite these encouraging results, several limitations remain. First, the learned image-quality branch is still supervised by generic IQA signals, which do not perfectly align with downstream product recognizability. Second, the current fusion stage relies on structured prompting and heuristic conflict handling, leaving room for more principled confidence calibration and multimodal alignment. Third, although the proposed system is designed for video scenarios, it does not yet explicitly model longer-range temporal dynamics beyond frame selection. Future work may therefore explore product-oriented quality supervision, lighter yet more expressive multi-frame reasoning, stronger temporal modeling, and broader validation on larger and more diverse livestream benchmarks.

In summary, the proposed "keyframe prioritization + multimodal LLM fusion" framework achieves consistent improvements in Perfect Match Rate, Product Name Recognition Accuracy, and Semantic Similarity, while maintaining a deployment-oriented computational profile. We believe the study provides both a strong empirical baseline and a useful design perspective for future work on recognition-oriented frame selection and multimodal decision making in real-world video understanding systems.

Author Contributions: Conceptualization, Y.Z., J.S. and W.S.; methodology, Y.Z., J.S. and W.S.; software, Y.Z.; validation, Y.Z., J.S. and W.S.; formal analysis, Y.Z. and W.S.; investigation, Y.Z., J.S. and W.S.; resources, J.S. and W.S.; data curation, Y.Z. and J.S.; writing—original draft preparation, Y.Z.; writing—review and editing, J.S. and W.S.; visualization, Y.Z.; supervision, J.S. and W.S.; project administration, W.S.; funding acquisition, J.S. and W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Plan Project of Zhejiang Provincial Administration for Market Regulation ("Research on the Intelligent Definition System of CCC Products Empowered by AI in the Context of Artificial Intelligence"), Grant No. LY2026018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and proprietary restrictions.

 doi:10.20944/preprints202605.0313.v1

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. During the preparation of this manuscript, the authors used ChatGPT and Gemini for language editing and grammatical corrections. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Wangjingshe E-commerce Research Center. The Release of "2024 China Live E-commerce Market Data Report". https://www.100ec.cn/detail--6649904.html, 2025. Accessed: 2025-06-09.
2. Transparency Market Research. Livestream E-commerce Market Size & Industry Share to 2034. Technical report, Transparency Market Research, 2025.
3. Pe, P. TikTok Shop GMV in 2024 Surpassed US$30 Billion. https://thelowdown.momentum.asia/tiktok-shop-gmv-in-2024-surpassed-us30-billion, 2025. Accessed: 2026-02-02.
4. Wangjingshe. TikTok E-commerce French Station Accelerates Expansion, Transaction Volume Soars Seven-fold in Half a Year. https://fgw.sz.gov.cn/ztzl/qtztzl/szscjmyjjfzzhfwpt/hwtz/sjal/content/post_12527556.html, 2025. Accessed: 2026-02-02.
5. Sellercraft. TikTok Shop vs Shopee GMV Trends in Southeast Asia (2023–2025). https://sellercraft.co/tiktok-shop-vs-shopee-gmv-trends-in-southeast-asia-2023-2025-unpacking-the-e-commerce-showdown/, 2025.
6. Yang, W.; Chen, Y.; Li, Y.; Cheng, Y.; Liu, X.; Chen, Q.; Li, H. Cross-view Semantic Alignment for Livestreaming Product Recognition. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Oct 2023, pp. 13404–13413.
7. for Market Regulation, S.A. The Fifth Batch of Typical Cases in the Field of Live E-commerce. https://www.samr.gov.cn/xw/zj/art/2026/art_a7a1fb24ceac4ed9a4789161bfe49e0f.html, 2026.
8. TikTok Shop. TikTok Shop's latest Safety and IPR Reports: Focusing on security while growing globally. https://seller.tiktokglobalshop.com/business/us/newsroom/detail/10023362, 2025.
9. Amazon. 2024 Brand Protection Report: How Amazon Uses AI Innovations to Stop Fraud and Counterfeits. Technical report, Amazon, 2025.
10. OECD and EUIPO. Mapping Global Trade in Fakes 2025: Global Trends and Enforcement Challenges. Technical report, OECD, 2025.
11. Zhu, H.; Wei, H.; Li, B.; Yuan, X.; Kehtarnavaz, N. A Review of Video Object Detection: Datasets, Metrics and Methods. *Applied Sciences* **2020**, *10*, 7834. https://doi.org/10.3390/app10217834.
12. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2019, p. 6558.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
14. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2020.
15. Peng, Y.; Li, H.; Wu, P.; Zhang, Y.; Sun, X.; Wu, F. D-FINE: Redefine Regression Task in DETRs as Fine-Grained Distribution Refinement, 2024.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems* **2017**, *30*.
17. Kuhn, H.W. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* **1955**, *2*, 83–97.
18. Neubeck, A.; Van Gool, L. Efficient Non-Maximum Suppression. In Proceedings of the Proceedings of the International Conference on Pattern Recognition (ICPR). IEEE, 2006, pp. 850–855.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems* **2015**, *28*.
20. Weng, Z.; Meng, L.; Wang, R.; Wu, Z.; Jiang, Y.G. A Multimodal Framework for Video Ads Understanding, 2021.

 doi:10.20944/preprints202605.0313.v1

21. Gu, J.; Qin, T.; Chen, H. A Key Frame Extraction Method Based on MPEG-7 Color Features and Block Motion Information. *Journal of Guangxi University (Natural Science Edition)* **2010**, *35*, 310–314. https://doi.org/10.13624/j.cnki.issn.1001-7445.2010.02.029.
22. Cernekova, Z.; Pitas, I.; Nikou, C. Information Theory-Based Shot Cut/Fade Detection and Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology* **2006**, *16*, 82–91. https://doi.org/10.1109/TCSVT.2005.856896.
23. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* **2012**, *21*, 4695–4708. https://doi.org/10.1109/TIP.2012.2214050.
24. Kharchevnikova, A.; Savchenko, A.V. Efficient Video Face Recognition Based on Frame Selection and Quality Assessment. *PeerJ Computer Science* **2021**, *7*, e391. https://doi.org/10.7717/peerj-cs.391.
25. Park, J.; Lee, J.; Kim, I.J.; Sohn, K. SumGraph: Video Summarization via Recursive Graph Modeling, 2020.
26. Zhuang, Y.; Rui, Y.; Huang, T.S.; Mehrotra, S. Adaptive Key Frame Extraction Using Unsupervised Clustering. In Proceedings of the Proceedings of the International Conference on Image Processing (ICIP), 1998, pp. 866–870. https://doi.org/10.1109/ICIP.1998.723655.
27. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* **2013**, *20*, 209–212. https://doi.org/10.1109/LSP.2012.2227726.
28. Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3664–3673. https://doi.org/10.1109/CVPR42600.2020.00372.
29. Liu, X.; van de Weijer, J.; Bagdanov, A.D. RankIQA: Learning from Rankings for No-Reference Image Quality Assessment, 2017.
30. Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; Ma, K. Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective, 2023.
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision, 2021.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.
33. Golestaneh, S.A.; Dadsetan, S.; Kitani, K.M. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency, 2022.
34. Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; Yang, Y. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment, 2022.
35. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In Proceedings of the Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2023, pp. 19730–19742.
36. Cui, C.; Sun, T.; Lin, M.; Gao, T.; Zhang, Y.; Liu, J.; Wang, X.; Zhang, Z.; Zhou, C.; Liu, H.; et al. PaddleOCR 3.0 Technical Report, 2025.
37. An, K.; Chen, Q.; Deng, C.; Du, Z.; Gao, C.; Gao, Z.; Gu, Y.; He, T.; Hu, H.; Hu, K.; et al. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs, 2024.
38. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training, 2021.