Know-How's: A Multi-modal Agent System for Converting Manufacturing Videos into Standard Operating Procedures

Jaegwang Sim
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
jksim1833@ajou.ac.kr

Jaemin Yoo
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
woals1211@ajou.ac.kr

Mumin Chun
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
kidgood1@ajou.ac.kr

Eungyeol Lee
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
comangs@ajou.ac.kr

Dongyun Sung
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
dy2514@ajou.ac.kr

Jinwook Choi
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
shs480025@ajou.ac.kr

Gitae Koh
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
northspear@ajou.ac.kr

Yongjin Kwon*
dept. of Industrial Engineering
Ajou University
Suwon, Republic of Korea
yk73@ajou.ac.kr

***Abstract*— The preservation of tacit knowledge in the manufacturing sector is becoming increasingly critical due to the rapid aging of the skilled workforce and the subsequent risk of skill loss. However, traditional knowledge transfer methods face limitations in capturing the embodied nuances of physical craftsmanship. In this paper, we propose a multi-modal AI framework designed to automatically transform unstructured manufacturing videos into standardized digital manuals, referred to as "Know-how Heritage." Unlike conventional video-to-text approaches, our system integrates Range of Motion (ROM) and Hand-Object Interaction (HOI) analysis to extract quantitative data on ergonomic postures and fine-grained tool manipulations. These multi-modal insights are orchestrated by Large Language Models (LLMs) to generate context-aware Standard Operating Procedures (SOPs). Large-scale experiments involving 3000 iterations on a shoe manufacturing dataset demonstrate that the proposed framework significantly outperforms baseline models, achieving a high Factuality score of 0.936 (G-Eval, 0-1 scale) and demonstrating superior structural coherence. This study provides a sustainable solution for digitizing industrial expertise, effectively bridging the gap between experts and the future workforce.**

Keywords—Tacit Knowledge Preservation, Multi-modal AI Framework, Range of Motion, Hand-Object Interaction, Large Language Model, Industrial Expertise Digitization

I. INTRODUCTION

The preservation and revitalization of the manufacturing sector stands out as a pressing challenge facing many economies today. In particular, for manufacturing powerhouses such as South Korea and Japan, the persistent aging of the industrial workforce has emerged as a critical societal issue. The demographic transition towards an aged society is a critical global phenomenon, posing significant threats to economic sustainability. In particular, Japan and South Korea are experiencing this shift at an unprecedented pace compared to other advanced economies. Their research indicates that Japan, already a super-aged society, is projected to see its elderly population reach 30.9% by 2040. Similarly, South Korea faces an even faster aging rate, with its working-age population expected to shrink by more than 25% between 2019 and 2040[1] However, this structural shift in the workforce precipitates a crisis beyond a mere labor shortage: the potential loss of high-skilled technologies, specifically "Tacit Knowledge", which forms the foundation of manufacturing competitiveness. While Polanyi [2] defined tacit knowledge as intrinsic knowledge difficult to articulate, Nonaka and Takeuchi [3] identified the 'externalization' process which is converting tacit into explicit knowledge as the most critical yet challenging aspect of knowledge management. According to Zhao, [4] tacit knowledge is deeply embedded in the experience and practical routines of individual skilled workers. While it provides a unique competitive advantage that is difficult to imitate, its un-coded nature makes it inherently difficult to explicitly transfer or share with others. Particularly in sophisticated processes such as New Product Development (NPD), the transfer of implicit know-how becomes a decisive factor governing product quality and success.

In the current manufacturing landscape, this core asset is disappearing without being digitized alongside the retirement of skilled workers, and traditional apprenticeship-based education methods alone face limitations in effectively preserving and inheriting this knowledge.

On the other hand, in the era of Large Language Models (LLMs), recent advancements offer a promising solution to this challenge. Hadi et al. [5] highlighted that the sophisticated natural language processing capabilities of LLMs demonstrate exceptional performance in interpreting unstructured data. Building on these capabilities, the application of LLMs is expanding into diverse domains through natural language generation. In particular, the recent emergence of Vision-Language Models (VLMs) has revolutionized this field. The Gemini Team [6] demonstrated that these multimodal models can process high-dimensional video information and rapidly

generate structured linguistic descriptions. This advancement enables the automated conversion of procedural video data into interpretable text, offering a direct pathway to digitize the tacit knowledge embedded in visual demonstrations.

In this study, we propose a multi-modal AI framework designed to automatically transform the tacit knowledge of skilled workers into a standardized digital manual, referred to as "Know-how Heritage." The proposed system consists of a three-phase pipeline: Data Preprocessing, Multi-modal Tacit Knowledge Analysis, and Manual Generation. By integrating computer vision models such as MediaPipe and SAM2 with the reasoning capabilities of the Google Gemini API. Our framework applies ergonomics principles, specifically through Range of Motion (ROM) and Hand-Object Interaction (HOI) analysis to systematically extract and synthesize behavioral, interactive, and verbal data into explicit knowledge.

Fig. 1. Visualization of multi-modal analysis results on a skilled worker's video. (Left) Range of Motion (ROM) analysis tracking upper body joints to quantify working posture. (Right) Hand-Object Interaction (HOI) analysis detecting hand landmarks and the Region of Interest (ROI) for the target object.

II. Related Work

A. LLM-based Tacit Knowledge Extraction

In the domain of knowledge management, the potential of Large Language Models (LLMs) to externalize implicit knowledge has gained significant attention. Zuin et al. [7] empirically verified the feasibility of employing LLM-based agents to interactively extract and document fragmented tacit knowledge within organizational contexts. Their study utilized an epidemic model simulation to demonstrate how an agent could reconstruct dataset descriptions by aggregating partial information distributed across a network of employees, proving that LLMs can recover lost knowledge without accessing a single central expert.

However, their approach primarily addresses declarative knowledge retrieval through conversational interactions in a synthetic environment. It does not address the embodied and procedural nature of tacit knowledge prevalent in the manufacturing sector, such as specific physical postures or fine-grained tool manipulations. To address this limitation, leveraging multi-modal capabilities becomes essential. A comprehensive survey by Tang et al. [8] categorizes recent advancements in "Vid-LLMs" and identifies the "Video Analyzer LMM" framework as a key paradigm. In this architecture, specialized external tools first extract structured information, which is then processed by the LLM for high-level reasoning. Tang et al. highlight that this modular approach effectively utilizes the reasoning power of LLMs for spatiotemporal understanding, enabling the interpretation of complex interactions.

Aligning with this multi-modal paradigm, our framework focuses on the direct and automated transformation of unstructured video and audio data captured from real-world skilled workers into standardized procedural manuals. Unlike Zuin et al.'s text-centric method, we adopt a "Video Analyzer LLM" structure where MediaPipe and SAM2 serve as specialized data analyzers to extract industrial procedural knowledge before the generating manual using LLMs, extending the scope of digitization from organizational memory to physical craftsmanship.

B. Automated Evaluation and LLM-as-a-Judge

Evaluating the quality of content generated by LLMs traditionally relies on human annotation, which is costly and lacks scalability. To address this challenge, recent studies have explored the "LLM-as-a-Judge" paradigm, employing LLMs to assess generated outputs. Miao et al. [9] introduced the Self-Check framework, demonstrating that LLMs possess the capability to zero-shot verify their own step-by-step reasoning processes. Their research highlights that this self-reflection paradigm enables models to identify logical errors and critique their own outputs without external supervision, thereby significantly improving reliability. Inspired by this self-verification capability, our framework incorporates an automated evaluation pipeline within the LLMOps system. Instead of relying solely on manual inspection, we utilize the LLM as an objective judge to rigorously assess the generated "Know-how Heritage" manuals against the raw multi-modal data.

C. Hybrid Evaluation Metrics

To ensure a robust and objective assessment of the generated manuals, recent research advocates for a multi-dimensional evaluation strategy. Following the methodology validated in tacit knowledge research [7], we adopt a hybrid evaluation approach that combines model-based reasoning with traditional lexical metrics. First, leveraging the self-verification capability established by Miao et al. [9], we employ G-Eval [10] as a primary metric. Liu et al. [10] demonstrated that G-Eval, which utilizes chain-of-thought (CoT) reasoning, achieves a higher correlation with human judgment than traditional metrics by evaluating semantic coherence and factuality. Second, to complement the probabilistic nature of LLM-based evaluation, we utilize METEOR [11]. Unlike simple overlap metrics, METEOR considers synonymy and stemming, providing a robust measure of lexical similarity against ground truth data. This dual-metric approach ensures a balanced evaluation between semantic accuracy (G-Eval) and lexical precision (METEOR).

III. Proposed Approach

Documenting the tacit knowledge of skilled workers in the manufacturing sector presents a unique challenge: their expertise is often embedded in physical nuances that are difficult to articulate verbally or capture through simple video recording-analyzing process. Traditional archiving methods merely store visual data, failing to extract the underlying "know-how" essential for effective skill transfer.

To address this, we propose a multi-modal AI framework designed not just for digital archiving but for ergonomics and operational knowledge extraction. Our approach goes beyond surface-level observation by integrating Range of Motion

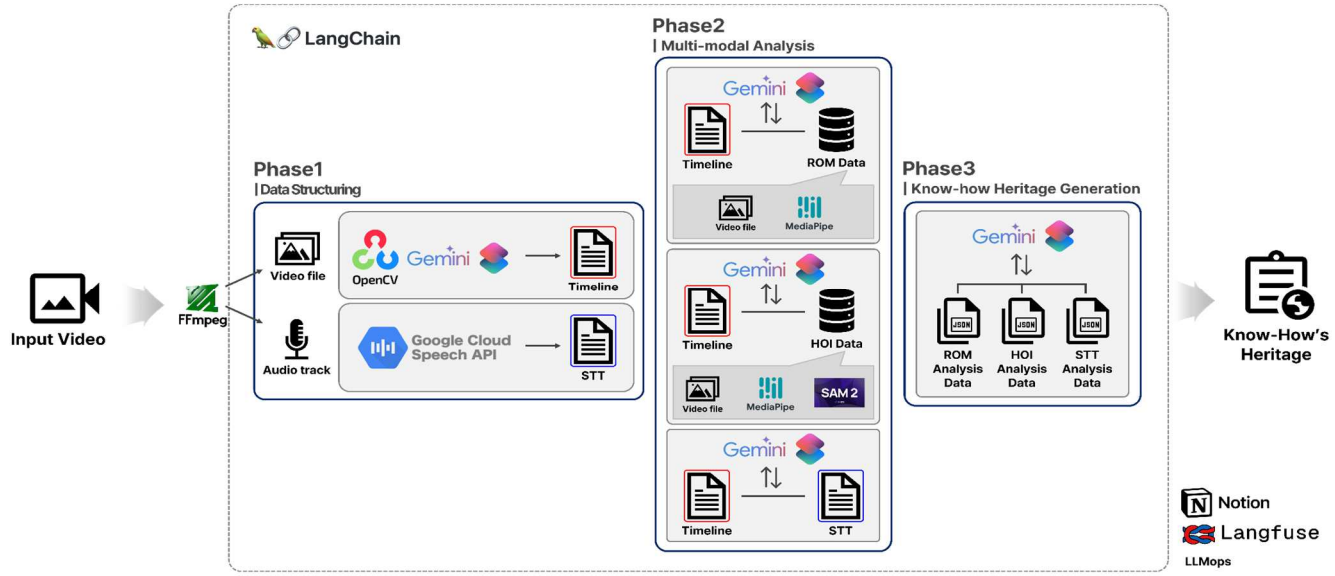

Fig. 2. The overall architecture of the proposed multi-modal AI framework. The pipeline consists of three phases: Phase 1 structures raw video and audio data using STT and timeline analysis. Phase 2 extracts tacit knowledge through Range of Motion (ROM) analysis via MediaPipe and Hand-Object Interaction (HOI) analysis via SAM2. Phase 3 synthesizes these multi-modal insights using the Google Gemini API to generate the standardized "Know-how Heritage" manual.

(ROM) analysis to quantify ergonomics efficiency and Hand-Object Interaction (HOI) analysis to decode the subtle, often unconscious, interactions between the worker and their tools.

As illustrated in Fig. 2, the proposed system is structured as a multi-layered pipeline consisting of three phases: Data Preprocessing, Multi-modal Tacit Knowledge Analysis, and Know-how Heritage Generation. By orchestrating these modules via the Google Gemini API, our framework systematically transforms unstructured video data into a standardized, context-aware manual, enabling the preservation of industrial heritage in a format that is actionable for beginners.

Phase 1: Data Structuring & Segmentation

The first phase focuses on structuring continuous raw video into analyzable units. We utilize FFmpeg to separate audio and video tracks from the input footage and employ the Google Cloud Speech API to convert speech data into text (STT). Simultaneously, OpenCV is used to decompose the video into frames, while a LangChain with Google Gemini API series based timeline analyzer automatically segments the entire process into meaningful operational units (e.g., preparation, cutting, finishing). This process transforms long-sequence video data into discrete, analyzable modules with JSON data files.

Phase 2: Multi-modal Tacit Knowledge Analysis

The second phase represents the core of our framework, extracting quantitative data from the embodied know-how of skilled workers. Beyond simple action recognition, we introduce Range of Motion (ROM) and Hand-Object Interaction (HOI) analysis to capture the subtle nuances that determine work quality. After the ROM and HOI modules extract quantitative data based on the interactions between the worker and the workpiece, the LangChain-integrated Google Gemini API synthesizes this information to generate a standardized JSON data file.

- **Range of Motion (ROM) Analysis**: Using MediaPipe Pose Estimation, we track the coordinates and angles of the worker's major joints in real-time. The primary value of ROM analysis lies in the numerical archiving of skilled working postures. By securing data on stable body angles and optimal working radii, unskilled workers can quantitatively compare their postures with the ideal model. This visualizes kinesthetic nuances often missed in apprenticeship education, thereby accelerating the learning curve and facilitating rapid skill acquisition.
- **Hand-Object Interaction (HOI) Analysis**: By combining MediaPipe Hands and the Segment Anything Model 2 (SAM2), we analyze the fine-grained interactions between hands and tools. Experienced workers unconsciously perform complex manipulations such as specific gripping methods, force distribution, and minute finger movements, which are difficult to articulate verbally. HOI analysis documents these unobtrusive skills as objective data. For instance, it detects changes in grip pressure or finger positioning during specific processes, converting what was once abstract tacit proficiency into explicit knowledge.

Phase 3: Know-how Heritage Generation

In the final phase, the structured timeline data, physical data (ROM), interaction data (HOI), and verbal explanations (Voice) extracted from Phases 1 and 2 are synthesized. This multi-modal data is fed into the Google Gemini 2.5 Flash API to generate a Markdown format Standard Operating Procedure (SOP) that is accessible and easy for beginners to follow. The resulting output is not merely a sequence of text but a comprehensive, formatted "Know-how Heritage." It is structured with distinct sections complemented by a ‘Master’s Tip’ segment that encapsulates implicit know-how. This layout directly correlates with our multi-modal analysis, where ROM data informs the 'Body' section and HOI data details the 'Hand' section, accompanied by key snapshot

images ensuring the permanent preservation of industrial intellectual assets.

Fig. 3. Sample frames from the real-world manufacturing datasets collected for this study. The dataset covers diverse domains requiring different types of tacit knowledge: (a) Shoe manufacturing and repair, (b) Manual welding, (c) Bicycle repair, and (d) Tailoring.

IV. Experiments

To validate the proposed framework, we constructed a comprehensive video dataset covering four distinct industrial sectors: shoe manufacturing, welding, bicycle repair, and tailoring, as shown in Fig. 3. The dataset consists of video footage capturing the detailed tasks of skilled workers. Among these, we selected the 'Shoe Manufacturing and Repair' dataset for our primary experiments. For a fair comparison, we utilized Google Gemini 2.5 Flash API as the backbone Large Language Model (LLM) for both the baseline and the proposed approach.

Following the methodology validated in recent tacit knowledge archiving research [7], we adopted a hybrid evaluation strategy combining semantic and lexical metrics.

- **G-Eval (Semantic Accuracy):** We employed G-Eval [10] to assess consistency with actual actions and compliance with the SOP format. Leveraging the self-verification capability of LLMs, G-Eval provides a 0-1 scale score that correlates highly with human judgment.
- **METEOR (Lexical Precision):** To complement model-based evaluation, METEOR[11] was used to measure the lexical overlap between the generated text and the ground truth data. The ground truth was rigorously constructed based on key technical specifications and know-how explicitly identified through interviews with skilled experts. This metric ensures that key technical terms were accurately preserved.

To validate the robustness and scalability of the proposed framework, we conducted a large-scale evaluation involving 3,000 iterations. Table 1 summarizes the comparative performance between the baseline (direct video input) and our proposed method.

TABLE I. Comparative Performance Analysis on Shoe Manufacturing Dataset (n=3,000)

Method	**Model**	**G-Eval (0-1)** Mean (SD)	**METEOR** Mean (SD)
Baseline	Gemini 2.5 Flash	0.411 (± 0.11)	0.152 (± 0.01)
Proposed (Ours)		0.911 (± 0.09)	0.282 (± 0.03)
Improvement	-	+121.6%	+85.5%

As presented in Table 1, the proposed framework achieved a G-Eval score of 0.911, demonstrating a substantial improvement over the baseline (0.411). This indicates that our ROM/HOI-based approach generates manuals with significantly higher factuality and structural coherence. Furthermore, the METEOR score of 0.282 (vs. 0.152) confirms that our method more accurately preserves key technical terminology and expert instructions.

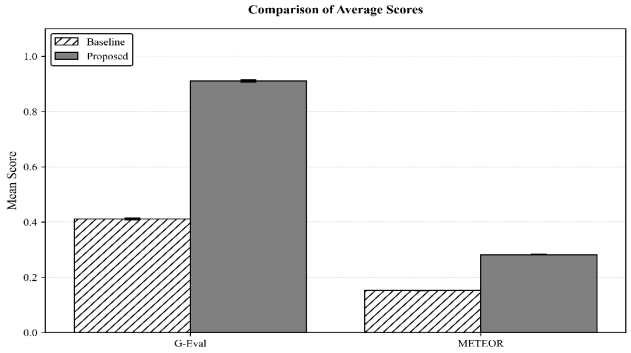

Fig. 4. Quantitative comparison of manual quality between the baseline (left) and the proposed framework (right) over n =3000 iterations, Average scores for Strcuture Complience and Factuality.

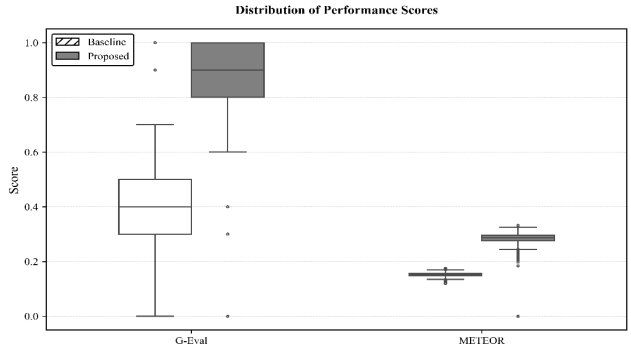

Fig. 5. Quantitative comparison of manual quality between the baseline (left) and the proposed framework (right) over n =3000 iterations. Score distribution density plots demonstrates consistency and reliability

Fig. 4 and 5 visualizes the comparative results of the large-scale experiment (n=3,000). The bar charts in the top row clearly indicate that the proposed framework outperforms the baseline in average scores for both structure compliance and factuality. More importantly, the density plots in the bottom row reveal a critical difference in system stability. The baseline (left) exhibits a broad and scattered score distribution, implying that the quality of the output fluctuates significantly depending on the input video complexity. In contrast, the proposed method (right) shows a sharp, narrow peak in the

high-score region. This signifies that our framework robustly maintains high-quality generation regardless of variations in the input data.

V. Discussion

After conducting extensive experiments with 3,000 iterations of the proposed multi-modal framework, we identified several critical insights regarding the system's technical efficacy and broader implications. The most significant finding is the decoupling of output quality from model size. Our results demonstrate that the quality of procedural documentation depends fundamentally on the granularity of structured input data, using with our framework rather than the parameter size of the LLM. This model-agnostic efficiency implies that industrial applications can achieve high-fidelity results using cost-effective models (e.g., Gemini Flash) by investing in robust pre-processing pipelines, making the solution viable for SMEs with limited computational resources.

Beyond these technical metrics, the utility of our framework extends to establishing a scalable and inclusive knowledge ecosystem. The text-based modular architecture allows for seamless integration with translation APIs, enabling the instant conversion of manuals into workers' native languages, thereby dismantling language barriers for the growing foreign workforce. Furthermore, the fine-grained coordinates from ROM and interaction details from HOI serve as high-quality ground truth data for Large Action Models (LAMs). This positions our system as a robust "Data Generator for Embodied AI," facilitating the transfer of human skills to industrial robots through imitation learning.

While promising, this framework presents challenges that warrant further investigation. First, the HOI analysis currently relies on a semi-automated process requiring human supervision for SAM 2 prompting. Future research will aim to automate this via open-vocabulary object detection models (e.g., Grounding DINO). Second, high scores in automated metrics do not guarantee practical usefulness for training. Subsequent studies will focus on validating the learning effectiveness through experiments with actual beginners.

VI. Conclusion

In this paper, we proposed a multi-modal AI framework designed to prevent the loss of industrial tacit knowledge by transforming it into a standardized "Know-how Heritage." By integrating Range of Motion and Hand-Object Interaction analysis with the reasoning capabilities of Large Language Models, our system successfully digitizes the embodied expertise of skilled workers into explicit, actionable manuals.

Our extensive experiments confirmed that the proposed framework significantly outperforms baseline methods in both factuality and structural compliance, while achieving substantial cost-efficiency. However, we acknowledge that the current system relies on semi-automated processes for object segmentation and lacks long-term pedagogical validation with human subjects. Future work will address these limitations by integrating fully autonomous vision models and conducting longitudinal user studies to verify educational effectiveness.

References

[1] J.-W. Lee, E. Song, and D. W. Kwak, “Aging labor, ICT capital, and productivity in Japan and Korea,” *J. Jpn. Int. Econ.*, vol. 58, p. 101095, Dec. 2020, doi: 10.1016/j.jjie.2020.101095.

[2] M. Polanyi and A. Sen, *The Tacit Dimension*. University of Chicago Press, 2009.

[3] I. Nonaka, “The Knowledge-Creating Company,” *Harvard Business Review*. Accessed: Nov. 29, 2025. [Online]. Available: https://hbr.org/2007/07/the-knowledge-creating-company

[4] Y. Zhao, “Tacit Knowledge Transfer from Manufacturing Firms to Suppliers in New Product Development: A Study of Suppliers,” *Int. J. Inf. Educ. Technol.*, vol. 3, no. 5, 2013.

[5] M. U. Hadi *et al.*, “A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage,” July 10, 2023. doi: 10.36227/techrxiv.23589741.v1.

[6] G. Team *et al.*, “Gemini: A Family of Highly Capable Multimodal Models,” May 09, 2025, *arXiv*: arXiv:2312.11805. doi: 10.48550/arXiv.2312.11805.

[7] G. Zuin, S. Mastelini, T. Loures, and A. Veloso, “Leveraging Large Language Models for Tacit Knowledge Discovery in Organizational Contexts,” July 04, 2025, *arXiv*: arXiv:2507.03811. doi: 10.48550/arXiv.2507.03811.

[8] Y. Tang *et al.*, “Video Understanding with Large Language Models: A Survey,” *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2025, doi: 10.1109/TCSVT.2025.3566695.

[9] N. Miao, Y. W. Teh, and T. Rainforth, “SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning,” Oct. 05, 2023, *arXiv*: arXiv:2308.00436. doi: 10.48550/arXiv.2308.00436.

[10] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment,” May 23, 2023, *arXiv*: arXiv:2303.16634. doi: 10.48550/arXiv.2303.16634.

[11] A. Lavie and A. Agarwal, “METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments,” in *Proceedings of the Second Workshop on Statistical Machine Translation*, C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz, Eds., Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 228–231. Accessed: Nov. 30, 2025. [Online]. Available: https://aclanthology.org/W07-0734/