

AUTOARABIC: A Three-Stage Framework for Localizing Video-Text Retrieval Benchmarks

Mohamed Eltahir¹
Taha Alshatiri¹

Osamah Sarraj¹
Mohammed Khurd¹
Tanveer Hussain²

Abdulrahman Alfrihidi¹
Mohammed Bremoo¹

¹ King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

² Department of Computer Science, Edge Hill University, Ormskirk, England

{mohamed.hamid@kaust.edu.sa, osamah.sarraj@gmail.com, frihidimany@gmail.com,
tahaalshatiri@gmail.com, mohamedalawi211@gmail.com, mohabremoo@gmail.com
hussaint@edgehill.ac.uk}

Abstract

Video-to-text and text-to-video retrieval are dominated by English benchmarks (e.g. DiDeMo, MSR-VTT) and recent multilingual corpora (e.g. RUDDER), yet Arabic remains underserved, lacking localized evaluation metrics. We introduce a three-stage framework, AUTOARABIC, utilizing state-of-the-art large language models (LLMs) to translate non-Arabic benchmarks into Modern Standard Arabic, reducing the manual revision required by nearly fourfold. The framework incorporates an error detection module that automatically flags potential translation errors with 97% accuracy. Applying the framework to DiDeMo, a video retrieval benchmark produces DiDeMo-AR, an Arabic variant with 40,144 fluent Arabic descriptions. An analysis of the translation errors is provided and organized into an insightful taxonomy to guide future Arabic localization efforts. We train a CLIP-style baseline with identical hyperparameters on the Arabic and English variants of the benchmark, finding a moderate performance gap ($\Delta \approx 3$ pp at Recall@1), indicating that Arabic localization preserves benchmark difficulty. We evaluate three post-editing budgets (zero/ flagged-only/ full) and find that performance improves *monotonically* with more post-editing, while the raw LLM output (zero-budget) remains *usable*. To ensure reproducibility to other languages, we made the code available at <https://github.com/Tahaalshatiri/AutoArabic>.

1 Introduction

The exponential growth of online video has created an urgent demand for accurate retrieval systems that can find relevant moments within long streams of visual content. On YouTube alone,



Figure 1: A sample of English captions and their MSA translations for three moments in the same video.

more than 500 hours of video are uploaded every minute (Shepherd, 2025).

Over the past decade, the research community has released a flood of English-centric benchmarks like DiDeMo (Anne Hendricks et al., 2017), MSR-VTT (Xu et al., 2016), the bilingual VATEX (Wang et al., 2019) and the multilingual RUDDER (Dabral et al., 2021).

Although these benchmarks have become standard for text-to-video and video-to-text retrieval, all of them completely omit Arabic. Subsequently, Arab researchers are forced to evaluate their retrieval models on English data, literally translated data, or private translations. This slows progress in Arabic multimodal research and questions the reproducibility of their results.

Our work helps fill this gap with a three-stage Large Language Models (LLMs) framework that localizes any non-Arabic retrieval benchmark into Modern Standard Arabic (MSA) with minimal human effort. The framework (i) uses a large language model to translate captions into Modern

Table 1: **Video-text retrieval benchmarks.** This table highlights a *language gap*: existing retrieval benchmarks are almost entirely English (with limited Chinese) and lack Arabic coverage. To our knowledge, only our DiDeMo-AR offers Modern Standard Arabic captions. "Moment-level" ✓ indicates that the dataset provides temporally-localized descriptions (segment boundaries).

Dataset	#Videos	Clip Len.	Languages	Moment-level	Arabic?
MSR-VTT (Xu et al., 2016)	10,000	15s	EN	✗	✗
VATEX (Wang et al., 2019)	41,250	10s	EN / ZH	✗	✗
DiDeMo (Anne Hendricks et al., 2017)	10,464	30s	EN	✓	✗
LSMDC (Rohrbach et al., 2015)	118,081	4-5s	EN	✗	✗
ActivityNet (Caba Heilbron et al., 2015)	19,994	120s	EN	✓	✗
RUDDER (Dabral et al., 2021)	100 k / lang.	5-10s	EN / ZH / FR / DE / RU	✗	✗
DiDeMo-AR	10,464	30s	AR	✓	✓

Table 2: **Arabic corpora with different modalities** (non-retrieval). This highlights a *task gap*: prior corpora focus on speech, sentiment, or QA and do not provide videotext retrieval benchmarks. DiDeMo-AR is the first publicly released Arabic dataset dedicated to retrieval.

Dataset	Modality	Primary Task	Size / Hours	Retrieval?
AmdSaEr (Haouhat et al., 2023)	Video + Audio + Text	Multimodal Sentiment	540 clips	✗
MGB-2 (Ali et al., 2016)	Audio + Subtitles	ASR (broadcast MSA)	~1200 h	✗
MASC (Al-Fetyani et al., 2023)	Audio	ASR (speech corpus)	~1000 h	✗
GALE Arabic (Glenn et al., 2017)	Audio + Text	ASR/MT (news/talk)	multi-year	✗
ArabicaQA (Abdallah et al., 2024)	Text	QA / Dense Retrieval	92k Q/A	✗
ANAD (Elnagar and Gouza, 2020)	Audio	Speech Emotion Rec.	1,700 utt.	✗
AVSD-Arabic (Elhaj and Abdulla, 2021)	Video + Audio	Lip-reading	1,100 vids	✗
DiDeMo-AR (ours)	Video + Text	Video Retrieval	40,144 caps	✓

Standard Arabic, (ii) utilizes a second LLM to automatically flag lexical, grammatical, and formatting errors, and (iii) sends only flagged samples to expert annotators for final verification. The workflow has been applied to the Distinct Describable Moments corpus (DiDeMo), resulting in **DiDeMo-AR**, the first Arabic video retrieval benchmark, consisting of 10,464 videos and 40,144 fluent Arabic descriptions. We further contribute the first systematic taxonomy of LLM translation errors for Arabic benchmark creation, intended as a reusable checklist for future translation efforts.

To ensure that localization preserves the original benchmark’s difficulty, we finetune a *Contrastive Language-Image Pre-training* (CLIP) baseline (Radford et al., 2021) that uses a *Vision Transformer* (ViT-B/16 and ViT-B/32) image encoder (Dosovitskiy et al., 2020) and a *Masked and Permuted Pre-training* (MPNet) text encoder (Song et al., 2020), optimized with the symmetric InfoNCE contrastive loss (van den Oord et al., 2018), on both the English and Arabic variants of DiDeMo. Although Arabic has a complex word structure, the model shows only a ≈ 3 -point drop in Recall@1 (R@1, higher is better). This result suggests that LLM-based translation, combined with light expert correction, can preserve benchmark difficulty without requiring language-

specific pre-training.

We believe this workflow, benchmark, and error analysis will help guide future Arabic benchmark localization research.

2 Background & Related Work

Early attempts to translate multimodal datasets relied either on direct machine translation of English captions or on small teams of human annotators. The MSVD-Indonesian corpus (Hendria, 2023), for example, was created by translating the original MSVD sentences into Indonesian with Google Translate and then finetuning a CLIP baseline. VATEX offers English-Chinese captions produced by human experts, but no Arabic version, and its captions are sentence-level rather than moment-level (Wang et al., 2019). RUDDER combines Google-translated captions with expert annotations and adds five additional languages, yet still omits Arabic entirely (Dabral et al., 2021). None of these projects publishes a detailed taxonomy of translation errors, so their contributions remain dataset-specific and provide little guidance for researchers who intend to localize new benchmarks.

Table 1 lists the retrieval benchmarks that have driven progress during the last decade. Corpora such as MSR-VTT (Xu et al., 2016) and LSMDC (Rohrbach et al., 2015) are clip-based and En-

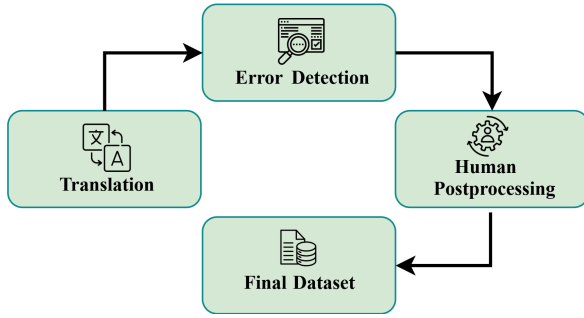


Figure 2: AUTOARABIC three-stage localization workflow: translation, error detection, and human post-editing.

English only. DiDeMo (Anne Hendricks et al., 2017) introduced moment-level ground-truth in $\sim 10k$ unedited Flickr videos, followed by ActivityNet Captions, which applies the same idea to long YouTube clips (Caba Heilbron et al., 2015).

Table 1 highlights a simple fact: not one public retrieval benchmark offers Modern Standard Arabic (MSA) captions, and only two (DiDeMo, ActivityNet) provide moment-level ground truth.

Looking into Arabic multimodal benchmarks, it can be seen that such benchmarks exist but they target tasks very different from retrieval. MGB-3 focuses on broadcast speech and automatic speech recognition (Ali et al., 2017). MASC provides more than 1,000 hours of YouTube audio for large-scale ASR experiments, again without video captions (Al-Fetyani et al., 2023). AmdSaEr utilizes short YouTube clips for sentiment and emotion recognition (Haouhat et al., 2023). Large text corpora such as ArabicaQA push reading comprehension research forward (Abdallah et al., 2024), yet contain no video. Table 2 summarizes the information from these datasets. To the best of our knowledge, **DiDeMo-AR** is therefore the first publicly released benchmark that pairs Arabic sentences with temporally grounded video moments.

3 The AUTOARABIC Framework

Figure 2 shows AUTOARABIC, a three-stage framework that can turn any English video-text benchmark into Modern Standard Arabic (MSA). In this section we describe the framework in general terms. Its output for DiDeMo is analyzed in the next sections.

First, every English caption is sent to Gemini 2.0 Flash (Cloud, 2025) with this prompt:

"You will receive an English sentence that serves as a caption for a short video clip.

Your task is to translate this caption into Modern Standard Arabic while ensuring that the translation remains suitable and appropriate as a caption.
The English caption: {caption}
Arabic caption:"

Gemini is run with temperature=0.7 and top-p=1.0. Next, each Arabic output is processed by GPT-4o (OpenAI, 2025) for automatic error detection, tagging six categories: lexical, literal, hallucination, tense_shift, loanword, and diacritics (summarized in Table 3).

Finally, captions flagged by the detector are reviewed by five native-speaker annotators. Although the framework supports selective post-editing, we performed a full revision in this study, where annotators reviewed every caption rather than only the flagged ones. We compared the error detection performance of the LLM against that of the annotators and found that the LLM successfully identified over 97% of the actual mistakes.

Using these reviewed captions, we evaluated caption quality under *three post-editing budgets*: (i) Raw LLM output (zero), (ii) Fix only LLM-flagged (few), and (iii) Fix all (full). Results show that performance improves monotonically with greater post-editing (zero $\rightarrow$ few $\rightarrow$ full), while the raw LLM output remains usable.

It is worth mentioning that the framework is provider-agnostic: the prompting, validation, and post-processing steps do not depend on a specific API and can be run with open or proprietary LLMs. In this paper, we used high-performing commercial models to maximize one-time localization quality.

Additionally, diacritics themselves are *not* errors; *inconsistency* across samples is. We intentionally did not constrain diacritics in the translation prompt to observe natural model behavior, then enforced uniformity post hoc via deterministic stripping.

4 The DiDeMo-AR Dataset

The Distinct Describable Moments (DiDeMo) dataset (Anne Hendricks et al., 2017) is one of the largest and most diverse datasets for the temporal localization of events in videos given natural language descriptions. The videos are collected from Flickr and each video is trimmed to a maximum of 30 seconds. The videos in the dataset are divided into 5-second segments to reduce the complexity

Table 3: Error categories identified by the automated detector and addressed through manual post-editing.

Error Type	Definition	Example (English / Arabic)
<i>Lexical</i>	Selection of uncommon or overly formal words instead of familiar alternatives.	EN: <i>first time we see an otter swim by</i> AR-poor: هذه أول مرة نرى فيها قضاة تسبح. AR-improved: هذه أول مرة نرى ثعلب الماء يسبح.
<i>Literal</i>	Word-for-word structural translation that produces unnatural Arabic phrasing.	EN: <i>The man raises onto his knees to crawl.</i> AR-poor: يرفع الرجل جذعه ليستند على ركبتيه زحفاً. AR-improved: ينفض الرجل على ركبتيه ليزحف.
<i>Hallucination</i>	Addition of content not present in the original English text.	EN: <i>The girl starts speaking.</i> AR-poor: الفتاة تبدأ بالتحدث باللغة العربية. AR-improved: الفتاة تبدأ بالتحدث.
<i>Tense Shift</i>	Incorrect temporal rendering of present actions in past tense.	EN: <i>Person in black exits frame to left.</i> AR-poor: يخرج الشخص ذو اللباس الأسود من المشهد نحو اليسار. AR-improved: يخرج الشخص ذو اللباس الأسود من المشهد نحو اليسار.
<i>Loanword</i>	Inconsistent use of transliterated terms versus established Arabic equivalents.	EN: <i>The camera zooms up on the players.</i> AR-poor: تقترب الكاميرا بالتكبير على اللاعبين. AR-improved: تقترب آلة التصوير بالتكبير على اللاعبين.
<i>Diacritics</i>	Inconsistent application of diacritical marks across words and captions.	EN: <i>The gentleman puts his left arm under his right arm.</i> AR-poor: يضع الرجل ذراعه اليسرى تحت ذراعه اليمنى. AR-improved: يضع الرجل ذراعه اليسرى تحت ذراعه اليمنى.

of annotation. The dataset is split into training, validation and test sets containing 8,395, 1,065 and 1,004 videos respectively. The dataset contains a total of 26,892 moments and one moment could be associated with descriptions from multiple annotators. The total number of captions in DiDeMo is 40,144. The descriptions in DiDeMo dataset are detailed and contain camera movement, temporal transition indicators, and activities. Moreover, the descriptions in DiDeMo are verified so that each description refers to a single moment.

Applying the translation framework to DiDeMo yields **DiDeMo-AR** with the same 10,464 videos and 26,892 moments, but now 40,144 fluent MSA captions. Arabic captions are slightly shorter, 5.6 words on average versus 7.5 in English. Figure 3 plots the word-per-caption distribution for both languages on the top, while Figure 4 visualizes the most frequent content words. It can be seen that the most common words in English also appear in the Arabic figure with nearly the same size, indicating consistent translation and semantic map-

ping across languages.

Table 4 reports unique n -gram and POS counts. While Arabic and English share a similar 1-gram vocabulary count, the counts diverge as we move to longer n -gram. Regarding POS tokens, Arabic shows a smaller set of distinct POS tokens compared to English. Achieving performance close to the English baseline with a smaller lexical set shows the concise expressive power of Arabic.

During manual revision, we logged the errors found in every caption. Their distribution is shown in Table 5, where error rate denotes the percentage of captions containing ≥ 1 instance of the category (totals can exceed 100% because a caption may contain multiple categories). The most frequent issue is inconsistent use of diacritics (some captions contain full diacritics while others have none) accounting for 27.8% of the entire dataset. Loanword handling ranks second (12.7%), followed by tense shifts (3.4%). Literal translations, rare lexical choices, and hallucinations together occur in fewer than 5% of captions.

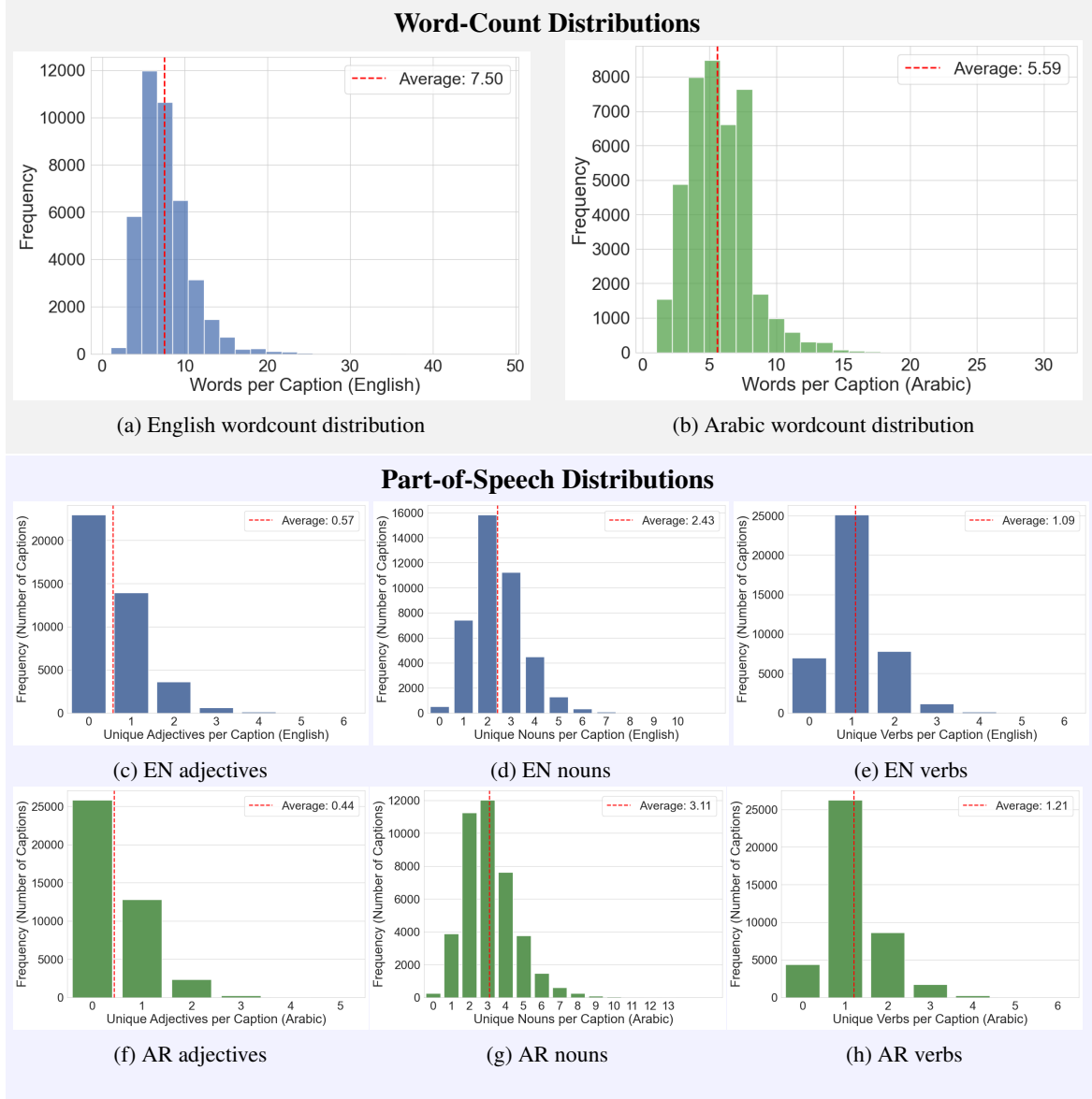


Figure 3: **Top:** Wordcount distributions per caption for English (left) and Arabic (right) in DiDeMo vs. DiDeMo-AR. **Middle:** Distributions of unique adjectives, nouns, and verbs per caption in English (DiDeMo). **Bottom:** Same distributions for Arabic (DiDeMo-AR).

Table 4: Unique n -grams and POS-tag counts in DiDeMo vs. DiDeMo-AR.

Language	1-gram	2-gram	3-gram	4-gram
English	5,358	67,698	140,387	163,841
Arabic	5,205	75,904	151,943	176,369
POS	verbs	nouns	adj.	adv.
English	1,320	3,605	891	333
Arabic	1,145	2,822	713	17

Table 5: **Top:** exclusive single-error rates on the DiDeMo-AR dataset. **Bottom:** distribution of captions that shows multiple error types simultaneously.

Error Type	%
Diacritics	27.8
Loanwords	12.7
Literal / weak phrasing	5.0
Tense shift	3.4
Hallucination	1.8
Total error rate (overlapped)	41.7
Overlap Type	%
Loanword + Diacritics	7.1
Tense shift + Diacritics	1.6
Tense shift + Loanword	0.4
Tense + Loan + Diac.	0.1

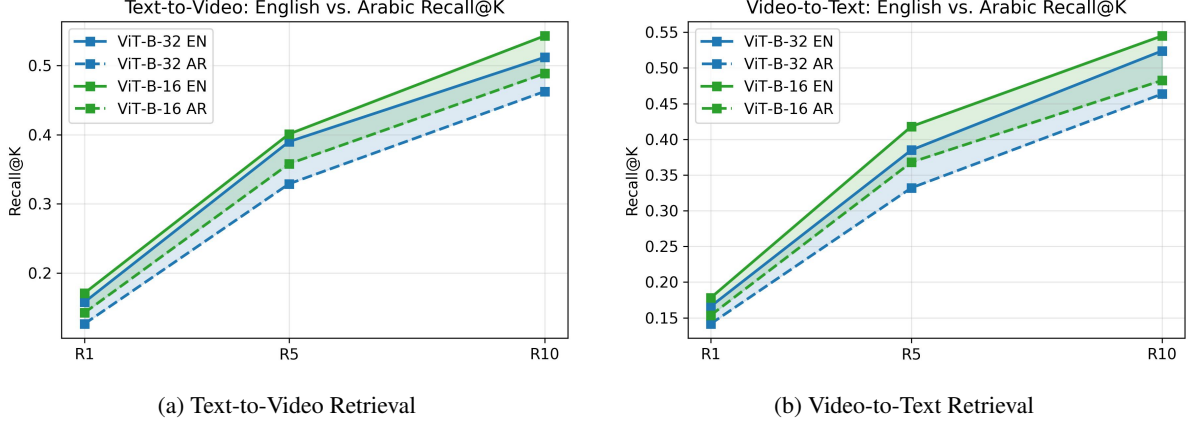


Figure 5: English vs. Arabic performance comparison in text-video and video-text retrieval (Recall@K).

Table 6: Text-to-Video retrieval performance on DiDeMo test split. Δ : the Arabic-English performance gap.

Model	Lang.	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MedR $\downarrow$	MeanR $\downarrow$
ViT-B-32 + MPNet	EN	0.158	0.390	0.512	10	48.2
	AR	0.127 (Δ -0.031)	0.329 (Δ -0.061)	0.463 (Δ -0.049)	13 (Δ +3)	55.7 (Δ +7.5)
ViT-B-16 + MPNet	EN	0.171	0.401	0.543	8	45.9
	AR	0.143 (Δ -0.028)	0.358 (Δ -0.043)	0.489 (Δ -0.055)	11 (Δ +3)	50.6 (Δ +4.7)

Table 7: Video-to-Text retrieval performance on DiDeMo test split. Δ : the Arabic-English performance gap.

Model	Lang.	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MedR $\downarrow$	MeanR $\downarrow$
ViT-B-32 + MPNet	EN	0.166	0.385	0.524	9	48.3
	AR	0.142 (Δ -0.024)	0.332 (Δ -0.053)	0.464 (Δ -0.060)	13 (Δ +4)	54.3 (Δ +6.0)
ViT-B-16 + MPNet	EN	0.178	0.418	0.545	8	44.9
	AR	0.154 (Δ -0.025)	0.368 (Δ -0.050)	0.483 (Δ -0.062)	11 (Δ +3)	49.8 (Δ +4.9)

5 Experiments & Results

5.1 Setup & Baselines

We fine-tune two CLIP backbones **ViT-B/32** and **ViT-B/16**, while *freezing* the vision tower and updating only a 256-d projection head. The text branch is paraphrase-multilingual-mpnet-base-v2 (768d; 110 M parameters). Training follows a symmetric InfoNCE loss, batch size 64, AdamW ($\text{lr} = 1\text{e-}4$, weight-decay $1\text{e-}2$) and runs for six epochs on one A100-80 GB. Input videos are down-sampled to eight uniformly spaced frames (224×224). We train identical scripts on the original English captions and on the new Arabic set, so any gap is purely linguistic. Our CLIP baseline is deliberately lightweight. Its role is to verify that the Arabic variant remains comparably difficult, not to exhaustively benchmark Arabic video-retrieval models.

5.2 Overall Retrieval Scores

Tables 6 and 7 report Recall@K, Median Rank, and Mean Rank on the DiDeMo test split. Despite Arabic captions being 25% shorter, the absolute drop is small: $\Delta\text{R@1} < 3$ pp for both ViT backbones in *text-to-video* and *video-to-text* directions. Median rank increases by three to four positions on average, but still stays below 15.

Figure 5 overlays English and Arabic curves. The shaded area highlights the gap. It never exceeds 0.07 at R@10. This shows that performance gaps remain nearly parallel across R@1, 5, 10.

Using the fully post-edited Arabic captions, a frozen CLIP backbone recovers 85-90% of its English Recall@10. This confirms that *metric localization* using our framework preserves benchmark difficulty without extra Arabic pre-training, with most of the English retrieval strength transferring directly to Arabic.

Table 8: Text-to-Video retrieval across post-editing levels on DiDeMo-AR.

Model	Post-Editing	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
ViT-B-16 + MPNet	Raw (zero)	0.1196	0.3230	0.4676	13.0	55.9
	Flagged-only (few)	0.1316	0.3121	0.4556	12.0	55.3
	Fix all (full)	0.1426	0.3579	0.4885	11.0	50.6
ViT-B-32 + MPNet	Raw (zero)	0.1176	0.3270	0.4636	13.0	55.2
	Flagged-only (few)	0.1157	0.3310	0.4646	13.0	54.9
	Fix all (full)	0.1266	0.3290	0.4626	13.0	55.7

Table 9: Video-to-Text retrieval across post-editing levels on DiDeMo-AR.

Model	Post-Editing	R@1 ↑	R@5 ↑	R@10 ↑	MedR ↓	MeanR ↓
ViT-B-16 + MPNet	Raw (zero)	0.1306	0.3519	0.4835	11.0	53.2
	Flagged-only (few)	0.1236	0.3500	0.4726	12.0	54.2
	Fix all (full)	0.1535	0.3679	0.4826	11.0	49.8
ViT-B-32 + MPNet	Raw (zero)	0.1296	0.3420	0.4646	12.0	54.6
	Flagged-only (few)	0.1286	0.3450	0.4646	13.0	54.4
	Fix all (full)	0.1416	0.3320	0.4636	13.0	54.3

5.3 Effect of Post-Editing Effort

To understand how human post-editing impacts retrieval performance, we evaluate three levels of manual correction on Arabic captions:

- **Raw (zero):** Direct LLM output without human intervention.
- **Flagged-only (few):** Corrections applied only to LLM-flagged captions.
- **Fix all (full):** Comprehensive manual review and correction of all captions.

Tables 8 and 9 show that even raw LLM translations achieve reasonable performance. However, increasing post-editing effort yields consistent improvements, with full correction typically providing ≈ 2 percentage points gains in R@1 across both retrieval directions.

Notably, if raw translations already work, then benchmark replication becomes language-agnostic, no per-language retraining or major human effort required, provided a capable translation LLM.

5.4 Automated Error-Flagging Quality

We evaluate the LLM-based error detector on our human-reviewed dataset. The automated system achieves strong agreement with human annotators: 97% accuracy and 91% F1-score (macro-averaged).

Table 10 shows the detector performs perfectly on diacritics and achieves high precision for hallucination detection. Tense shifting proves most

challenging (F1=0.80), reflecting the complexity of Arabic temporal expressions.

Table 10: Per-class precision, recall, and F1-score of the automated error-flagging system.

Class	Precision	Recall	F1
Diacritics	1.00	1.00	1.00
Hallucination/Literal	1.00	0.92	0.96
Loanword	0.91	0.82	0.86
No Error	0.93	0.97	0.95
Tense Shifting	0.77	0.84	0.80
Overall (macro-avg)	0.92	0.91	0.91

Limitations & Future Work

Our study takes a first step toward Arabic-centric video-text retrieval, but richer domains, dialects and modalities remain wide open for exploration.

Generalization. Our findings suggest that direct machine translation may enable language-agnostic benchmark replication without per-language retraining. Extending this beyond DiDeMo and MSA, across datasets, domains, and dialects, remains an open direction for future work.

Dataset Scope. DiDeMo-AR covers short clips (30 s) captured in real-world conditions. Long-form videos such as movies, lectures, or sports broadcasts are out of scope. Future work could localize MAD corpus (Soldan et al., 2022) or the LOVR benchmark (Cai et al., 2025), for example,

to MSA and dialects, giving researchers a benchmark for *long-video retrieval*.

Language Coverage. We focus on Modern Standard Arabic. Dialects, like: Egyptian, Gulf and Maghrebi, are still missing, yet they dominate social media videos (Guellil et al., 2021). A fruitful extension is to repeat the framework for *dialectal captions*.

Acknowledgments

We are grateful to the KAUST Academy for its generous support, and especially to Prof. Sultan Albarakati and Prof. Naeemullah Khan for providing the resources and guidance that made this work possible.

References

- Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049--2059.
- Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2023. Masc: Massive arabic speech corpus. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1006--1013. IEEE.
- Ahmed Ali, Peter Bell, James R. Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. [The mgb-2 challenge: Arabic multi-dialect broadcast media recognition](#). In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279--284, San Diego, CA, USA. IEEE.
- Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316--322. IEEE.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803--5812.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961--970.
- Qifeng Cai, Hao Liang, Hejun Dong, Meiyi Qiang, Ruichuan An, Zhaoyang Han, Zhengzhou Zhu, Bin Cui, and Wentao Zhang. 2025. Lovr: A benchmark for long video retrieval in multi-modal contexts. arXiv:2505.13928.
- Google Cloud. 2025. Gemini 2.0 flash on vertex ai: Low-latency multimodal generation. <https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/gemini>. Accessed June 2025.
- Rishabh Dabral, Ganesh Ramakrishnan, Preethi Jyothi, and 1 others. 2021. Rudder: A cross lingual video and text retrieval dataset. *arXiv preprint arXiv:2103.05457*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M. Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Salah Elhaj and Waleed Abdulla. 2021. Avsd-arabic: An audio-visual lip-reading dataset for modern standard arabic. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.
- Amal Elnagar and Ahmed Gouza. 2020. Anad: Arabic natural audio dataset for speech emotion recognition. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.
- Meghan Glenn, Haejoong Lee, Stephanie Strassel, and Kazuaki Maeda. 2017. [Gale phase 4 arabic broadcast conversation transcripts](#). LDC2017T12, Web Download.
- Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. [Arabic natural language processing: An overview](#). *Journal of King Saud University - Computer and Information Sciences*, 33(5):497--507.

- Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. 2023. Towards arabic multimodal dataset for sentiment analysis. In *2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pages 126--133. IEEE.
- Willy Fitra Hendria. 2023. Msvd-indonesian: A benchmark for multimodal video-text tasks in indonesian. *arXiv preprint arXiv:2306.11341*.
- OpenAI. 2025. GPT-4o: Openais omnimodal flagship model. <https://openai.com/blog/gpt-4o>. Accessed June 2025.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748--8763. PmLR.
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202--3212.
- Jack Shepherd. 2025. 23 essential youtube statistics you need to know in 2025. <https://thesocialshepherd.com/blog/youtube-statistics>. Updated June 11, 2025; accessed July 2, 2025.
- Francisco Soldan and 1 others. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kaitao Song, Xu Tan, Tianyang Zhang, Rui Wang, Liang Lu, Ada Lin, Qingyu Zhou, Lu Zhang, and Furu Wei. 2020. MPNet: Masked and Permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:1807.03748.
- Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuanfang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581--4591.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288--5296.