DuwatBench: Bridging Language and Visual Heritage through an Arabic Calligraphy Benchmark for Multimodal Understanding

Shubham Patle[1†] Sara Ghaboura[1†] Hania Tariq[2] Mohammad Usman Khan[3]
Omkar Thawakar[1] Rao Muhammad Anwer[1] Salman Khan[1,4]
[1]Mohamed bin Zayed University of AI, [2]NUCES, [3]NUST, [4]Australian National University
{shubham.patle, sara.ghaboura, omkar.thawakar}@mbzuai.ac.ae
https://mbzuai-oryx.github.io/DuwatBench/

Abstract

Arabic calligraphy represents one of the richest visual traditions of the Arabic language, blending linguistic meaning with artistic form. Although multimodal models have advanced across languages, their ability to process Arabic script, especially in artistic and stylized calligraphic forms, remains largely unexplored. To address this gap, we present DuwatBench, a benchmark of 1,272 curated samples containing about 1,475 unique words across 6 classical and modern calligraphic styles, each paired with sentence-level detection annotations. The dataset reflects real-world challenges in Arabic writing, such as complex stroke patterns, dense ligatures, and stylistic variations that often challenge standard text recognition systems. Using DuwatBench, we evaluated 13 leading Arabic and multilingual multimodal models and showed that while they perform well in clean text, they struggle with calligraphic variation, artistic distortions, and precise visual–text alignment. By publicly releasing DuwatBench and its annotations, we aim to advance culturally grounded multimodal research, foster fair inclusion of Arabic language and visual heritage in AI systems, and support continued progress in this area. Our dataset [1] and code[2] are publicly available.

1 Introduction

Arabic calligraphy sits at the meeting point of language and visual art. Letters are bound, stretched, and ornamented; diacritics and ligatures introduce significant variation in glyph shapes and baselines; style rules vary across schools such as Thuluth, Diwani, Naskh, and Kufic. These properties make calligraphy culturally and technically challenging for machine perception. Prior works ((Lorigo and Govindaraju, 2006), (Al-Salman and Alyahya, 2017)) have noted the added difficulty of Arabic script due to positional letter forms, dense ligaturing, and heavy use of diacritics, especially when written artistically rather than as plain text.

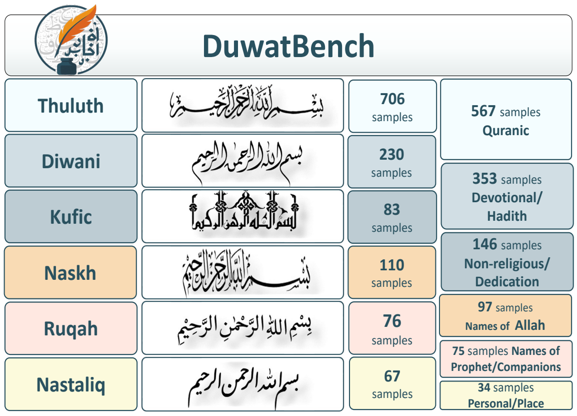

Figure 1: **DuwatBench Taxonomy and Distribution.** DuwatBench encompasses six principal Arabic calligraphy styles (Thuluth, Diwani, Kufic, Naskh, Ruq'ah, and Nasta'liq). The benchmark includes a distribution of collected samples ranging from 706 in Thuluth to 67 in Nasta'liq. Beyond style diversity, the taxonomy incorporates categories, including non-religious terms, Quranic words, devotional expressions and hadith, names of the Prophet and companions, names of Allah, and person or place names. This organization provides a structured basis for evaluating both the visual variation and semantic depth of Arabic calligraphy.

Despite growing interest in Arabic handwriting and manuscripts, existing datasets only partially address calligraphic use cases. Calliar (Alyafeai et al., 2022) provides pen-trace data across multiple text levels but lacks the textures and visual complexity of real artworks. HICMA (Ismail et al., 2023) includes real-world calligraphy images with sentence-level and style labels, but is limited to five scripts, focuses mainly on recognition, and lacks layout annotations as well as evaluation on large multimodal models (LMMs). Earlier works on style classification (Kaoudja, 2022) or letter-level corpora (Adam et al., 2017; Salamah and King, 2018) remain too narrow to support comprehensive multimodal assessment.

[1]https://github.com/mbzuai-oryx/DuwatBench
[2]https://huggingface.co/datasets/MBZUAI/DuwatBench
†Equal contribution.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5256–5269

Feature	Allaf	Adam	Sal&King	Kaoudja	Calliar	MOJ-DB	HICMA	DuwatBenach *(Ours)*
Real-world images	✓	✓	✓	✓	✗	✓	✓	✓
Style diversity (multiple styles)	✗	✓	✓	✓	✓	✗	✓	✓
Word / Sentence-level text	✓	✗	✗	✓	✓	✓	✓	✓
Full text transcriptions	✗	✗	✗	✗	✓	✓	✓	✓
Bounding boxes for detection	✗	✗	✗	✗	✗	✗	✗	✓
Complex artistic backgrounds	✓	✗	✗	✗	✗	✓	✓	✓
Publicly available	✗	✗	✗	✗	✓	✓	✓	✓

Table 1: **Comparison of existing Arabic calligraphy datasets.** Existing resources typically focus on isolated aspects such as limited style coverage, word- or sentence-level annotations, or constrained availability. In contrast, DuwatBench uniquely integrates multiple underrepresented dimensions like script diversity, word- and sentence-level transcriptions, bounding boxes for detection, and complex artistic backgrounds, while ensuring public accessibility. Allaf (Allaf and Al-Hmouz, 2016), Adam (Adam et al., 2017), Sal&King (Salamah and King, 2018), Kaoudja (Kaoudja, 2022), Calliar (Alyafeai et al., 2022), MOJ-DB(Zoizou et al., 2022), HICMA (Ismail et al., 2023).

Meanwhile, Arabic-capable multimodal models are advancing rapidly. Cross-lingual CLIP variants (Chen et al., 2023) and multilingual VLMs (Geigle et al., 2023) demonstrate strong retrieval and captioning capabilities across languages; Arabic-specific CLIP adaptations such as AraCLIP (Al-Barham et al., 2024) report clear gains on Arabic image–text benchmarks; and recent Arabic OCR and VLM systems (Bhatia et al., 2024; Wasfy et al., 2025; Heakl et al., 2025) show improved handling of diacritics and high-resolution text. However, none of these models have been evaluated on stylized Arabic calligraphy with complex artistic backgrounds and diverse style rules. A benchmark centered on calligraphy can thus reveal whether these systems can generalize to culturally grounded, visually demanding Arabic text, or still rely on surface-level cues.

We introduce DuwatBench, a 1.27K-sample benchmark for Arabic calligraphy designed to evaluate LMMs under realistic conditions. The dataset encompasses over 1,475 unique words drawn from both religious and cultural domains, including Quranic verses, devotional phrases, greetings, and poetic or proverbial expressions. It features 6 major calligraphic styles: Thuluth, Diwani, Naskh, Kufic, Nasta'liq, and the more modern Ruq'ah, capturing the visual diversity of Arabic writing across historical and contemporary contexts (see Figure 1). Each image is paired with transcriptions and detection-level annotations, enabling both recognition and localization analysis.

Unlike existing datasets that primarily target OCR-like tasks (Table 1), DuwatBench integrates real-world calligraphy with complex backgrounds, multiple scripts, full transcriptions, and bounding box annotations. Arabic calligraphy lies at the edge of visual–textual alignment: while meaning is preserved, artistic strokes, ligatures, and curved layouts often distort the patterns leveraged by models trained on clean Latin text. Evaluating multimodal models on DuwatBench therefore provides a rigorous test of whether Arabic or multilingual pretraining transfers to artistic calligraphy, distinguishing genuine understanding from pattern memorization.

Our contributions are threefold: (1) we release a curated benchmark of 1.27K samples spanning 6 calligraphic styles and 1,475 unique words, enriched with transcriptions and artistic contexts; (2) we introduce detection-level annotations for word localization, complementing prior classification- and OCR-based datasets; and (3) we establish systematic baselines on state-of-the-art Arabic and multilingual LMMs, revealing consistent failure cases such as style misinterpretation, sensitivity to diacritics and hamza, curved text alignment, and interference from background clutter.

DuwatBench bridges Arabic visual heritage and modern AI, enabling inclusive evaluation, style-aware modeling, and applications in preservation, education, and cultural search.

2 DuwatBench: The Arabic Calligraphy Dataset

2.1 Dataset Taxonomy

Our dataset corpus is structured along two main dimensions: textual categories and calligraphic styles (see Figure 1). On the content side, the largest portion consists of religious material, including Quranic verses, devotional invocations, and the Names of Allah, complemented by references to the Prophets and companions, as well as smaller portions of hadith and personal dedications. Non-religious expressions are also well

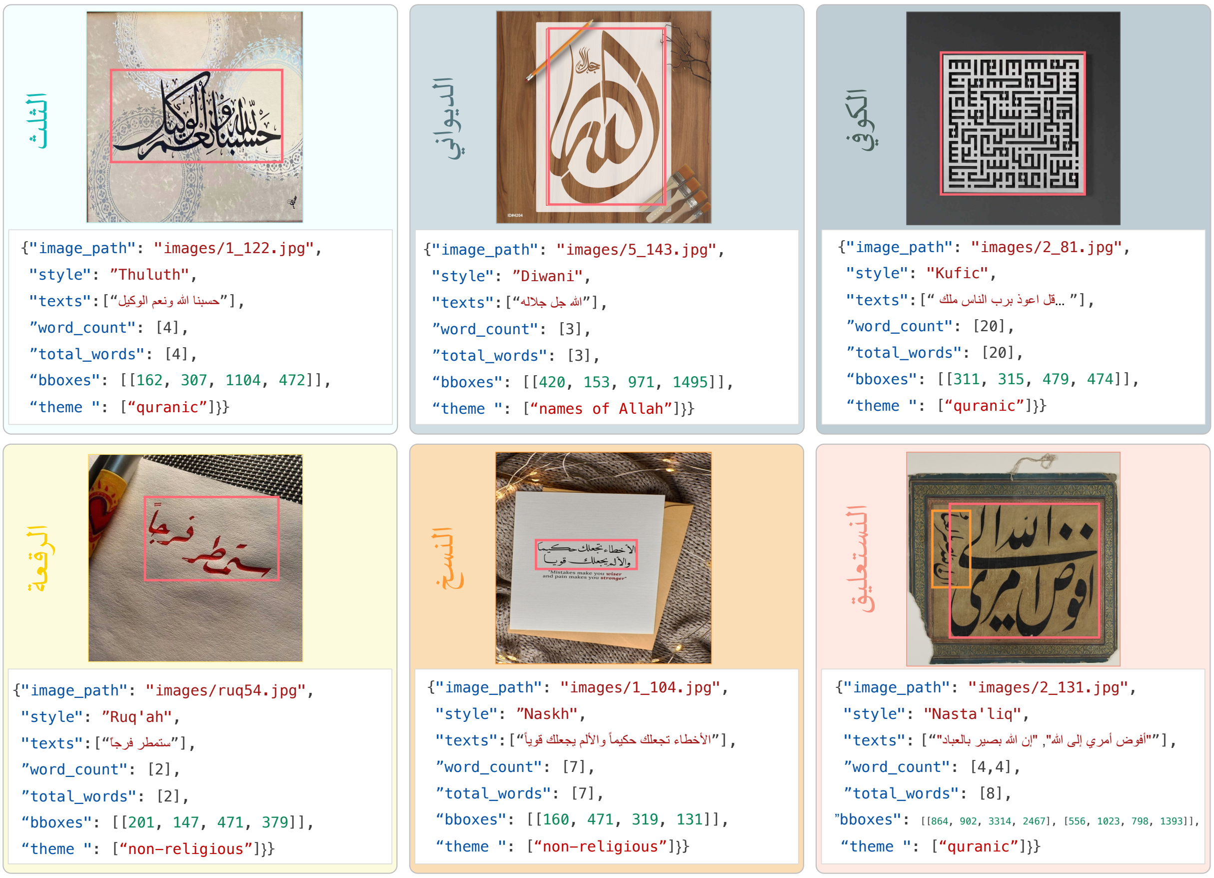

Figure 2: **Examples of DuwatBench Calligraphic Styles.** Representative samples from the six calligraphic styles in DuwatBench: Thuluth, Diwani, Kufic, Ruq'ah, Naskh, and Nasta'liq. Each entry displays the artwork with bounding box annotations, transcription, and metadata such as style, text content, theme, and word count. These examples highlight the dataset's diversity in structure, composition, and artistic context, spanning both religious and non-religious inscriptions.

represented, such as greetings, motifs, and cultural quotes common in decorative and public settings (see Figure 6). This balance ensures fair representation of both spiritual and cultural expressions in the dataset.

In terms of visual representation, the dataset covers 6 major calligraphic styles (Thulth, Diwani, Kufic, Nash, Ruq'ah, and Nasta'liq) with varying frequencies (see Figure 7). Further detailed statistical breakdown is provided in Appendix C.

2.2 Data Collection and Candidate Selection

As shown in Figure 3, the pipeline starts by sourcing candidate images from digital archives [3,4] and community repositories [5,6,7]. These candidates are screened for resolution, completeness, and the presence of authentic artistic backgrounds. Low-quality, blurred, or incomplete inscriptions are discarded at this stage, ensuring that only visually reliable samples move forward. We began with more than 2,950 samples, and after initial filtering and duplicate removal, the dataset was reduced to 1,285 high-quality instances, representing an overall reduction of approximately 57%.

[3]https://www.loc.gov/collections/
[4]https://digitalcollections.nypl.org/
[5]https://calligraphyqalam.com/
[6]https://freeislamiccalligraphy.com/
[7]https://www.pinterest.com/

2.3 Filtering, Transcription, & Normalization

Manual Transcription and Spatial Annotation. The retained samples are transcribed by 3 volunteer native Arabic speakers, who also generate bounding boxes using a free, open-source application[8] to capture the spatial alignment between words, characters, and stylistic forms. Since textual content varies across images, a single image may include multiple bounding boxes corresponding to distinct textual segments. Each sample is annotated with its transcription and style label, establishing a foundation for multimodal alignment between calligraphic form and semantic meaning as shown in Figure 4. To ensure annotation consistency, detailed guidelines were provided to annotators in advance (see

[8]https://www.makesense.ai

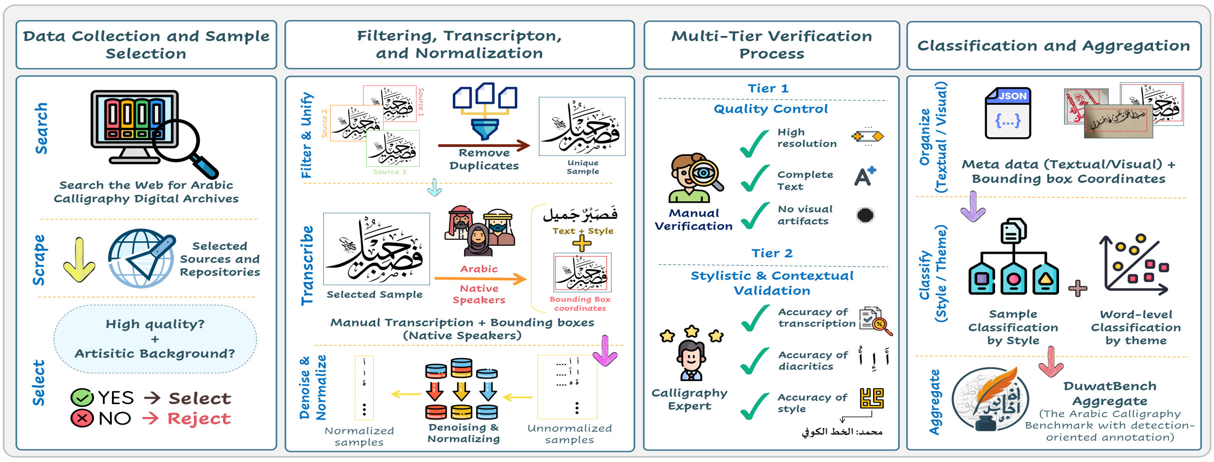

Figure 3: **Data collection and verification pipeline for DuwatBench.** The process begins with sourcing candidate images from digital archives and community repositories, followed by quality-based selection and manual transcription by native Arabic speakers. Each sample is denoised, normalized, and assigned bounding boxes with text–style and theme annotations. A multi-tier validation stage involving quality control, stylistic checks, and expert review ensures both textual accuracy and calligraphic fidelity. Final samples are classified and aggregated by style and thematic category, forming the curated DuwatBench corpus.

Table 5 in Appendix D).

Normalization and De-duplication. Then, samples undergo denoising and normalization to unify diacritics, letter variants like alef, hamza, ta'a marbouta, and encoding conventions to ensure that the corpus is both clean and standardized without compromising stylistic richness.

2.4 Multi-tier Verification

Quality assurance was conducted through a two-tier validation framework. Tier 1 addressed objective checks such as resolution, completeness, and artifact detection, while Tier 2 involved a calligraphy expert who verified transcription accuracy, diacritic integrity, and style and theme classification. After this process, the dataset was finalized with 1,272 high-quality samples. This layered approach combined systematic review with expert validation to ensure reliability and consistency of the annotations. More details on filtering criteria, verification steps, and annotation guidelines are provided in Appendix D.

2.5 Classification and Aggregation

Finally, validated samples are classified by script type and thematic category before being aggregated into the final benchmark. Each sample is accompanied by structured metadata (including transcription, bounding box coordinates, word counts, and style annotations) forming the DuwatBench corpus. This process ensures that the dataset captures both visual diversity and semantic richness, making it a robust resource for multimodal learning. Figure 2 provides representative examples, illustrating how each image is paired with its corresponding annotations.

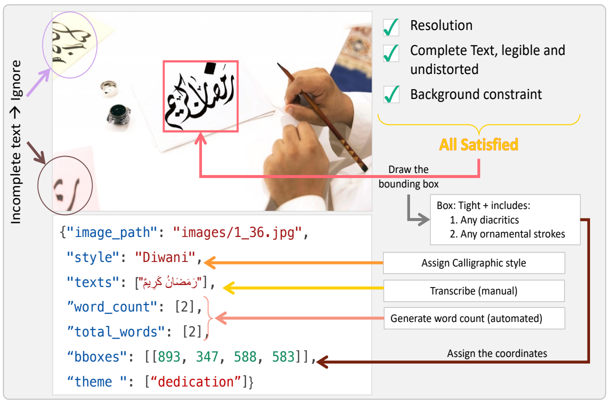

Figure 4: **Annotation Workflow in DuwatBench.** The figure illustrates the end-to-end annotation process followed in DuwatBench, from visual quality screening and bounding box drawing to transcription, style labeling, and word-count generation. Each stage ensures completeness, legibility, and consistent metadata alignment between text and visual form, supporting accurate multimodal evaluation.

3 Benchmark Evaluation

3.1 Evaluation Metric

We assess model performance with 5 complementary metrics: character error rate (CER), word error rate (WER), character F-score (chrF), ExactMatch, and normalized Levenshtein distance (NLD). CER captures the minimum number of character edits required to match the reference and is particularly important for Arabic calligraphy, where diacritics and ligatures can alter meaning. WER evaluates accuracy at the word level, offering a more interpretable measure for end-users. chrF computes

	Models	CER_mean ↓	WER_mean ↓	chrF↑	ExactMatch ↑	NLD_error↓
Open-source	Llava-v1.6-mistral-7b-hf (Liu et al., 2023)	0.9932	0.9998	9.1582	0.0000	0.9114
	EasyOCR (Kittinaradorn et al., 2024)	0.8538	0.9895	12.3016	0.0031	0.8163
	InternVL3-8B (Chen et al., 2024)	0.7588	0.8822	21.7461	0.0574	0.7132
	Qwen2.5-VL-7B (Team, 2025)	0.6453	0.7768	36.9703	0.1211	0.5984
	Qwen2.5-VL-72B-Instruct(Team, 2025)	0.5709	0.7039	43.9791	0.1761	0.5298
	Gemma-3-27B-IT (Team et al., 2025)	0.5556	0.6591	51.5330	0.2398	0.4741
	trocr-base-arabic-handwritten* (Microsoft, 2022)	0.9728	0.9998	1.7938	0.0000	0.9632
	MBZUAI/AIN* (Heakl et al., 2025)	0.5494	0.6912	42.6675	0.1895	0.5134
Closed-source	claude-sonnet-4.5 (Anthropic, 2025)	0.6494	0.7255	42.9660	0.2225	0.5599
	gemini-1.5-flash (Reid et al., 2024)	0.3933	0.5112	63.2790	0.3522	0.3659
	gemini-2.5-flash (AI, 2025)	0.3700	0.4478	71.8174	0.4167	0.3166
	gpt-4o-mini (OpenAI, 2024)	0.6039	0.7077	42.6717	0.2115	0.5351
	gpt-4o (OpenAI, 2024)	0.4766	0.5692	56.8510	0.3388	0.4245

Table 2: **Model Performance on DuwatBench.** Performance of open and closed-source models is reported using five complementary metrics: CER, WER, chrF, ExactMatch, and NLD. Results highlight clear differences across models, with larger multimodal models (gemini-2.5-flash and Gemma-3-27B-IT) showing stronger robustness, while others continue to face challenges on highly stylized calligraphic text. *Note:* * indicates Arabic-specific models. Values in blue denote the best performance, while underlined values indicate the second-best within each category (open or closed-source).

character n-gram F-scores, rewarding partial overlaps and providing robustness to tokenization errors and spelling variants common in stylized text. ExactMatch is the strictest metric, counting only perfect matches between prediction and reference. Finally, NLD normalizes edit distance by sequence length, situating itself between CER and WER in granularity and offering a balanced perspective on recognition errors across variable word sizes.
Together, these metrics provide a balanced evaluation framework, capturing fine-grained character and word errors (CER, WER), normalized string similarity (NLD), partial correctness under visual distortion (chrF), and strict full-sequence accuracy (ExactMatch).

3.2 Evaluation Settings

All metrics were computed using standard open-source implementations to ensure reproducibility and consistency. The *editdistance* and *Levenshtein* libraries were used for computing character- and word-level edit operations (CER, WER, NLD), while *sacrebleu* was employed for calculating chrF scores following machine translation evaluation standards. This setup provides a transparent and reproducible evaluation pipeline across all models.

To ensure fair comparison, we apply an Arabic-aware normalization pipeline before scoring. This includes Unicode normalization, removal of tatweel (_), and unification of character variants such as Alef forms (ا آ، إ، أ) and Alef Maqsurah (ى → ي). Such preprocessing follows standard practice in Arabic NLP tools (e.g., CAMeL Tools, Lucene Arabic normalizer) and avoids penalizing models for superficial script variants.

4 Results, Analysis, and Error Analysis

4.1 Quantitative Results

Table 2 presents the performance of open and closed-source models evaluated directly on full calligraphy images. The results reveal clear performance gaps, reflecting the models' varying capacity to interpret the artistic and structural complexity of Arabic calligraphy.

Among open-source systems, Gemma-3-27B-it achieved the best overall performance, recording the lowest CER (0.56) and WER (0.66) and the highest chrF (51.53) and ExactMatch (0.24) within this group. Followed by MBZUAI/AIN and Qwen2.5-VL-72B-Instruct, with balanced results across all metrics, indicating their stronger adaptation to the Arabic script. In contrast, traditional OCR baselines such as EasyOCR and trocr-base-arabic-handwritten performed poorly, struggling with the ornate ligatures and curved baselines typical of calligraphy. Similarly, general-purpose LMMs such as LLaVA-v1.6-Mistral-7B underperformed, suggesting that instruction tuning without explicit Arabic grounding is insufficient for stylized text recognition.

Among closed-source systems, Gemini-2.5-Flash achieves the strongest overall performance, with the highest chrF (71.82), the best ExactMatch

	Models	Kufic	Thuluth	Diwani	Naskh	Ruq'ah	Nasta'liq
Open-source	Llava-v1.6-mistral-7b-hf (Liu et al., 2023)	1.0000	0.9996	1.0000	1.0000	1.0000	1.0000
	EasyOCR (Kittinaradorn et al., 2024)	0.9895	0.9993	0.9990	0.9172	0.9757	0.9880
	InternVL3-8B (Chen et al., 2024)	0.9416	0.8584	0.8537	0.9260	0.9973	0.9438
	Qwen2.5-VL-7B (Team, 2025)	0.8942	0.7578	0.8184	0.6443	0.8324	0.8583
	Qwen2.5-VL-72B-Instruct (Team, 2025)	0.8658	0.6764	0.7640	0.5593	0.7613	0.7662
	Gemma-3-27B-IT (Team et al., 2025)	0.7802	0.6315	0.7326	0.5138	0.7571	0.6637
	trocr-base-arabic-handwritten* (Microsoft, 2022)	1.0000	0.9997	1.0000	0.9996	1.0000	1.0000
	MBZUAI/AIN* (Heakl et al., 2025)	0.7916	0.7036	0.7130	0.5367	0.6111	0.6916
Closed-source	claude-sonnet-4.5 (Anthropic, 2025)	0.8965	0.6724	0.8219	0.5859	0.8893	0.7659
	gemini-1.5-flash (Reid et al., 2024)	0.7212	0.4741	0.5783	0.4444	0.5445	0.5023
	gemini-2.5-flash (AI, 2025)	0.7067	0.3527	0.5698	0.4765	0.5817	0.5222
	gpt-4o-mini (OpenAI, 2024)	0.8790	0.6991	0.7825	0.5016	0.6908	0.6428
	gpt-4o (OpenAI, 2024)	0.8041	0.5540	0.6370	0.4189	0.5507	0.4434

Table 3: **WER Mean across Arabic calligraphy styles on DuwatBench.** WER scores are reported per script style (Kufic, Thuluth, Diwani, Naskh, Ruq'ah, Nasta'liq) for open and closed-source models. Lower values indicate better recognition accuracy. *Note:* * indicates Arabic-specific models. Values in blue denote the best performance, while underlined values indicate the second-best within each category (open or closed-source).

	Models	CER_mean ↓	WER_mean ↓	chrF↑	ExactMatch ↑	NLD_error↓
Open-source	Llava-v1.6-mistral-7b-hf (Liu et al., 2023)	0.9998	1.0000	7.0348	0.0000	0.9420
	EasyOCR (Kittinaradorn et al., 2024)	0.8935	0.9917	12.0570	0.0018	0.8330
	InternVL3-8B-Instruct (Chen et al., 2024)	0.7330	0.8566	22.6430	0.0957	0.7021
	Qwen2.5-VL-7B-Instruct (Team, 2025)	0.5957	0.7341	37.8491	0.1805	0.5602
	Qwen2.5-VL-72B-Instruct (Team, 2025)	0.5683	0.6898	42.9576	0.2201	0.5242
	Gemma-3-27B-IT (Team et al., 2025)	0.5494	0.6572	48.9358	0.2646	0.4707
	trocr-base-arabic-handwritten* (Microsoft, 2022)	0.9746	0.9997	1.9925	0.0000	0.9626
	MBZUAI/AIN* (Heakl et al., 2025)	0.5313	0.6692	43.9536	0.2233	0.4912
Closed-source	claude-sonnet-4.5 (Anthropic, 2025)	0.6896	0.7523	36.9575	0.2134	0.6076
	gemini-1.5-flash (Reid et al., 2024)	0.4110	0.5217	58.0771	0.3738	0.3823
	gemini-2.5-flash (AI, 2025)	0.3693	0.4531	66.5662	0.4488	0.3212
	gpt-4o-mini (OpenAI, 2024)	0.5975	0.6995	41.3582	0.2317	0.5290
	gpt-4o (OpenAI, 2024)	0.6107	0.7010	40.6754	0.2463	0.5437

Table 4: **Model Performance on DuwatBench with Bounding Boxes.** The table reports CER, WER, chrF, ExactMatch, and NLD for open- and closed-source models evaluated on DuwatBench using bounding box annotations. gemini-2.5-flash maintains the best overall performance across metrics, while within open-source models, MBZUAI/AIN achieves the lowest CER, outperforming Gemma-3-27B-IT. *Note:* * indicates Arabic-specific models. Values in blue denote the best performance, while underlined values indicate the second-best within each category (open or closed-source).

(0.42), and the lowest NLD error of (0.32), outperforming all baselines. Gemini-1.5-Flash ranks second, showing consistent robustness and surpassing GPT-4o and GPT-4o-mini in ExactMatch and NLD, while Claude-Sonnet-4.5 attains moderate performance.

Overall, these findings confirm that large, instruction-optimized multimodal models, particularly Gemini-2.5-Flash and Gemma-3-27B-IT, are more resilient to the structural irregularities and stylistic diversity of Arabic calligraphy than conventional OCR tools or smaller open models. This highlights the importance of large-scale multilingual pretraining and cross-modal grounding for robust Arabic visual text understanding.

Table 3 further reports WER across the six DuwatBench calligraphic styles. Models perform best on Naskh and Ruq'ah, particularly among open-source systems, where strokes are standardized and geometric distortion is limited. In contrast, ornate scripts such as Diwani and Thuluth remain challenging due to dense ligatures and fluid baselines. Nasta'liq exhibits variances, with some models achieving low error while others fail markedly, suggesting uneven exposure to this script during pretraining. Among open-source models, Gemma-3-27B-IT shows the most consistent cross-style performance, while MBZUAI/AIN records the lowest errors for Diwani and Ruq'ah, indicating familiarity with the two styles and modern Arabic handwriting. Within closed-source systems, Gemini-2.5-Flash, Gemini-1.5-Flash, and GPT-4o exhibit relatively stable performance across styles. Never-

Image											
Style	Diwani	Diwani	Diwani	Diwani	Thuluth	Thuluth	Ruqua	Kufic	Naskh	Naskh	Nasta'liq
Ground Truth	ما شاء الله	الرحيم	بسم الله	الحمد لله رب العالمين	بسم الله الرحمن الرحيم	فتوكل على الله	اللهم قلبًا يليق بلقائك.	فاصبر صبر جميلا	يا الله	أستغفر الله	خنده
Open-source											
llava-v1.6-mistral-7b-hf	،،،،-،،	،،،،،	،،،،	،،،،	،،،،	،،،	،،،	،،،،،	،""""،،-،،-،	الله	،،،،،
InternVL3-8B	الله	الله	الله	الله	الله أعلم	الله سميع بصير	الله يهدينا	الله	الله	استغفرالله	د ل ز
EasyOCR	ثمام | _ | م |، | ر	٧ | د			رللدالحخكن احمم | نمر	فتّكا كحللد	فادماً ب ابق بافّائك !0 د3_ 4.071ه	}	يا الل | ٥ط | ظ٥ ٥	الله | ستغفر	٥ | حلل
Qwen2.5-VL-7B	الله يحميك	الله أكبر	الله	الله	بسمة مريم بنت سليمان بن عقاص	فتوكة عَلى الله	أَمْرُ بَيْنَ يَدَائِكِ، قَلْبِي،	الجنة	يا الله	أَسْتَغْفِرُالله	خنده
Qwen2.5-VL-72B	الله	السلام عليكم	الله	الله	الله لا إله إلا	فتوك الله	اللهم فلياً بين بلقائك.	الله لا إله إلا	يَا الله	أستغفرالله	خنده
gemma-3-27b-it	الله محمد	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	يا حي يا قيوم	بسم الله الرحمن الرحيم	فتوكل على الله	اللهم قلباً بلين بلقائك.	يا عليُّ يا عظيمُ	يَا اللَّهُ	أستغفرالله	خنده
trocr-base-arabic-handwritten	.	لخر	.	درا | ط نا | ء بث نس	عندنانانا شيخ جاري الخيمة	.	.	.	.	في	.
MBZUAI/AIN	بسم الله	اللهم صل على محمد	بسم الله	الله هو أرحمهم	بسم الله الرحمن الرحيم	فتوكاً على الله	الله فليا يليق بالفائلك	الله هو الاكبر	يا الله	أستغفرالله	خذوه
Closed-source											
claude-sonnet-4.5	ماشاء الله	أريحوا قلوبكم	بسم الله	الإبداع	بسم الله الرحمن الرحيم	فتوكل على الله	أقلام تبحث عن كاتب.	طاب مساؤك	يا الله	استغفر الله	حنيد
gemini-1.5-flash	اللهُ شَهِيدٌ	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	بسم الله الرحمن الرحيم	فِي جَنَاحِ اللهِ	اللهم ولياً بين بلطافك.	بسم الله الرحمن الرحيم	يا الله	أستغفر الله	خالد
gemini-2.5-flash	ما شاء الله	الرحمن الرحيم | كلثوم	بسم الله	الرحمن الرحيم	بسم الله الرحمن الرحيم	فَتَوَكَّلْ عَلَى اللَّهِ	اللهم قلباً يلين بلقائك.	إِنَّا لِلَّهِ وَإِنَّا إِلَيْهِ رَاجِعُونَ	يا الله	أستغفر الله	خنده
gpt-4o-mini	ما شاء الله	الحمد لله	بسم الله	مبارك عليكم	بسم الله الرحمن الرحيم	فَوَكَّلْتُ عَلَى اللَّهِ	اللهم خَلِّ بَيْنَ يَدَيْكَ.	لا إله إلا الله	يَا اللَّهِ	أستغفر الله	خِذْهُ
gpt-4o	بسم الله	آسف | لا أستطيع المساعدة في ذلك	البسملة	أنا لا أستطيع تمييز النصوص داخل الصور | لكن إذا كنت بحاجة للمساعدة في أي شيء آخر | سأكون سعيدًا بمساعدتك	آسف | لا أستطيع نسخ النص من الصورة	فَتَوَكَّلْ عَلَى اللَّهِ	اللهم قلباً، يلين بلطفك.	آمَنُوا	آسف | لا أستطيع المساعدة في ذلك	آستغفر الله	لديك

Figure 5: **Qualitative Comparison of Model Outputs Across Calligraphic Styles** Examples from DuwatBench showing transcription quality across six calligraphic styles, evaluated using 13 open and closed-source OCR and vision-language models. Each column presents the model's predicted text aligned with the ground truth. The figure highlights challenges such as stroke density, curved baselines, and complex letter connections that complicate recognition, particularly in Thuluth, Diwani, and Kufic scripts. *Note:* Cells highlighted in green indicate correct predictions where diacritic variations are ignored; inaccurate or misplaced hamza marks are considered errors.

theless, occasional outliers, particularly on Naskh, Ruq'ah, and Nasta'liq, suggest variability in script-specific robustness, likely reflecting differences in script coverage during pretraining (additional results in Table 6, Appendix E).

These results reveal that model accuracy correlates strongly with style prevalence in pretraining data. Scripts such as Naskh and Ruq'ah, which dominate modern Arabic documents, yield the highest recognition rates, whereas historical and ornamental styles such as Kufic and Diwani remain the most error-prone.

To examine the impact of localization, we conducted ablation experiments in which models were evaluated using bounding box guidance that isolates calligraphic text regions. As shown in Table 4, this setting improves recognition accuracy for the majority of models, with particularly notable gains in ExactMatch and NLD, indicating more accurate holistic transcription when background clutter is reduced. Several open-source models, including Gemma-3-27B-IT, Qwen2.5-VL-72B-Instruct, and MBZUAI/AIN, show consistent improvements under localized input, suggesting that focused visual grounding better supports fine-grained character recognition in stylized scripts. Among closed-source systems, Gemini-2.5-Flash remains the strongest performer in both settings, maintaining robust accuracy with and without bounding boxes. In contrast, traditional OCR-based methods (EasyOCR, TrOCR) remain weak, underscoring their limited capacity to generalize beyond printed or lightly cursive Arabic text. While bounding box cropping mitigates background interference, cross-style imbalance persists, emphasizing the need for targeted data augmentation and style-aware pretraining.

In general, incorporating bounding box localization leads to consistent performance gains across most models, particularly enhancing recognition accuracy and alignment in visually complex calligraphic text. Nevertheless, fully addressing the challenges posed by Arabic calligraphy requires continued efforts toward improving coverage of underrepresented styles and reducing pretraining bias.

4.2 Error Analysis

Quantitative Error Analysis. Table 7 summarizes the performance of the models in the five metrics. The correlation analysis (see Figure 9 in Appendix F) reveals strong relationships between character and word errors ($\rho_{CER,WER} = 0.98$), as well as strong inverse correlations between normalized edit distance and semantic overlap (NLD–chrF = –0.99, NLD–ExactMatch = –0.95), indicating that lexical deviations closely align with losses in visual and semantic fidelity.

Closed-source models consistently outperform open-source counterparts, particularly in normalized edit distance ($p = 0.0112$) and exact-match accuracy ($p = 0.0029$); however, all systems struggle with complex scripts and long inscriptions. Performance degrades as word count and stylistic curvature increase, with the largest errors observed in Thuluth and Diwani samples, where ligatures, flourishes, and overlapping baselines amplify recognition difficulty. In contrast, models perform more reliably on shorter phrases and on Naskh and Ruq'ah texts, which exhibit more linear structure and consistent spacing.

Across the dataset, average ExactMatch remains below 0.18, underscoring the persistent challenge of faithfully transcribing stylized Arabic calligraphy even for advanced multimodal LLMs. These quantitative trends align with the qualitative analysis (Figure 5), where increased stylistic and visual complexity directly corresponds to recognition failures. Together, these results validate the proposed metric suite and provide a rigorous quantitative basis for cross-model performance comparison on DuwatBench. More details are provided in Appendix F.

Qualitative Error Analysis. Figure 5 visualizes representative outputs across six calligraphic styles, highlighting correct predictions in green. The results show substantial variation between models and styles. Among open-source systems, LLaVA-v1.6-Mistral-7B often produces incoherent Arabic, InternVL3-8B tends to overgenerate generic tokens like "الله", and EasyOCR exhibits fragmented outputs. Gemma-3-27B-IT achieves the most stable open-source performance, while Qwen2.5-VL variants perform best on simpler scripts like Naskh and Ruq'ah. MBZUAI/AIN performs competitively for its smaller scale, notably producing a correct Diwani recognition case, an area where most other open-source models fail.

In the closed-source category, Claude-Sonnet-4.5 and Gemini-2.5-Flash produce the most reliable transcriptions across Diwani and Thuluth, while GPT-4o-mini remains consistent, albeit slightly weaker, on Thuluth. GPT-4o frequently rejects input, reflecting stricter safety filtering.

Certain qualitative patterns also emerge. Models occasionally generate semantically plausible but visually mismatched outputs. For example, Gemma-3-27B-IT correctly recognized "بسم الله" but expanded it to the full Basmala "بسم الله الرحمن الرحيم," indicating cultural priors influencing prediction. Across models, a tendency to overpredict "الله" reflects strong cultural associations learned during training. Meanwhile, styles like Kufic and Diwani, which are geometrically rigid or ornate, remain the hardest to decode, while Naskh, Ruq'ah, and Nasta'liq achieve the highest recognition stability.

In summary, these findings highlight both linguistic and cultural biases in Arabic multimodal models and show that visual-textual grounding (rather than model size alone) is key to advancing calligraphy understanding.

5 Conclusion

In this work, we introduce DuwatBench, a benchmark for Arabic calligraphy comprising more than 1.27K curated samples that span various text categories and calligraphic styles. The dataset combines full transcriptions, style annotations, and detection-level information with authentic artistic backgrounds, offering a realistic testbed for Arabic LMMs. Our evaluation shows that, while existing systems handle standard text, they struggle with the stylistic variation and visual complexity of calligraphy. The benchmark highlights these limitations and encourages future research on script-aware modeling and culturally grounded AI, with applications in cultural preservation, education, and digital humanities.

6 Limitations and Societal Impact

Our proposed benchmark is designed as a focused and high-quality resource for evaluating Arabic calligraphy in multimodal models. Although the scale of the data set is smaller than that of large general-purpose Arabic corpora, it emphasizes diversity between text categories and calligraphic styles and ensures meaningful evaluation scenarios. Beyond technical contributions, the dataset highlights the cultural and artistic significance of Arabic calligraphy and supports research in areas such as cultural heritage preservation, digital archiving, and education. With this release, we aim to promote responsible and inclusive research practices that respect the cultural significance of the Arabic script while advancing the development of robust multimodal systems.

References

Kalthoum Adam, Somaya Al-Maadeed, and Ahmed Bouridane. 2017. based classification of arabic scripts style in ancient arabic manuscripts: Preliminary results. In *2017 1st international workshop on arabic script analysis and recognition (ASAR)*, pages 95–98. IEEE.

Google AI. 2025. Gemini 2.5 flash. Large language model (Preview), accessed May 20, 2025.

Muhammad Al-Barham, Imad Afyouni, Khalid Almubarak, Ashraf Elnagar, Ayad Turky, and Ibrahim Hashem. 2024. Araclip: Cross-lingual learning for effective arabic image retrieval. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 102–110.

AbdulMalik Al-Salman and Haifa Alyahya. 2017. Arabic online handwriting recognition: a survey. In *proceedings of the 1st international conference on internet of things and machine learning*, pages 1–4.

Wafa Alghallabi, Ritesh Thawkar, Sara Ghaboura, Ketan More, Omkar Thawakar, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Fann or flop: A multigenre, multiera benchmark for arabic poetry understanding in llms. *arXiv preprint arXiv:2505.18152*.

Siraj Reda Allaf and R Al-Hmouz. 2016. Automatic recognition of artistic arabic calligraphy types. *Journal of King Abdulaziz University*, 27(1):3–17.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. *arXiv preprint arXiv:2407.18129*.

Fakhraddin Alwajih, Samar Mohamed Magdy, Abdellah El Mekki, Omer Nacar, Youssef Nafea, Safaa Taher Abdelfadil, Abdulfattah Mohammed Yahya, Hamzah Luqman, Nada Almarwani, Samah Aloufi, et al. 2025a. Pearl: A multimodal culturally-aware arabic instruction dataset. *arXiv preprint arXiv:2505.21979*.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025b. Palmx 2025: The first shared task on benchmarking llms on arabic and islamic culture. *arXiv preprint arXiv:2509.02550*.

Zaid Alyafeai, Maged S Al-shaibani, Mustafa Ghaleb, and Yousif Ahmed Al-Wajih. 2022. Calliar: an online handwritten dataset for arabic calligraphy. *Neural Computing and Applications*, 34(23):20701–20713.

Anthropic. 2025. Claude-sonnet 4.5. `https://www.anthropic.com`. Large language model.

Farah Atif, Nursultan Askarbekuly, Kareem Darwish, and Monojit Choudhury. 2025. Sacred or synthetic? evaluating llm reliability and abstention for religious questions. *arXiv preprint arXiv:2508.08287*.

Gagan Bhatia, El Moatez Billah Nagoudi, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2024. Qalam: A multimodal llm for arabic optical character and handwriting recognition. *arXiv preprint arXiv:2407.13559*.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Mohamed Ezz, Mohamed A Sharaf, and Al-Amira A Hassan. 2019. Classification of arabic writing styles in ancient arabic manuscripts. *International Journal of Advanced Computer Science and Applications*, 10(10).

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad Shahbaz Khan, Salman H Khan, and Rao Muhammad Anwer. 2025a. Camel-bench: A comprehensive arabic lmm benchmark. *NAACL*.

Sara Ghaboura, Ketan More, Wafa Alghallabi, Omkar Thawakar, Jorma Laaksonen, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025b. Arb: A comprehensive arabic multimodal reasoning benchmark. *arXiv preprint arXiv:2505.17021*.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model.

Anis Ismail, Zena Kamel, and Reem Mahmoud. 2023. Hicma: the handwriting identification for calligraphy and manuscripts in arabic dataset. In *Proceedings of ArabicNLP 2023*, pages 24–32.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. *Preprint*, arXiv:2503.21910.

Zineb Kaoudja. 2022. *Arabic calligraphy style identification*. Ph.D. thesis, University of Kasdi Merbah Ouargla.

Zineb Kaoudja, Belal Khaldi, and Mohammed Lamine Kherfi. 2020. Arabic artistic script style identification using texture descriptors. In *2020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP)*, pages 113–118. IEEE.

Zineb Kaoudja, Mohammed Lamine Kherfi, and Belal Khaldi. 2021. A new computational method for arabic calligraphy style representation and classification. *Applied Sciences*, 11(11):4852.

Rakpong Kittinaradorn, Jaided AI Team, Wisuttida Wichitwong, Chan Kim, Danial Zakaria, Sumitkumar Sarda, Jeff Potter, Sam_S, Dakota Horstman, Ronald, Vijayabhaskar, Nina, DaeJeong Mun, Mijoo Kim, Amit Agarwal, Mejans, Vladimir Gurevich, Márton Tim, Tsogt Otgonbaatar, dependabot[bot], Karol Kucza, Loay, Korakot Chaovavanich, A2va, Abderrahim Mama, Vifly, TheDetective, Pray S, kfjt, and Wannaphong Phatthiyaphaibun. 2024. Jaidedai/EasyOCR. https://github.com/JaidedAI/EasyOCR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Liana M Lorigo and Venugopal Govindaraju. 2006. Offline arabic handwriting recognition: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):712–724.

Microsoft. 2022. Trocr-base-arabic-handwritten. `https://huggingface.co/microsoft/trocr-base-arabic-handwritten`. Transformer-based OCR model for Arabic handwritten text.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Mehreen Saeed, Adrian Chan, Anupam Mijar, Gerges Habchi, Carlos Younes, Chau-Wai Wong, Akram Khater, et al. 2024. Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition. *Advances in Neural Information Processing Systems*, 37:58525–58538.

Seetah AL Salamah and Ross King. 2018. Towards the machine reading of arabic calligraphy: a letters dataset and corresponding corpus of text. In *2018 IEEE 2nd international workshop on arabic and derived script analysis and recognition (ASAR)*, pages 19–23. IEEE.

Afnan Sumayli and Mohamed Alkaoud. Handwritten arabic calligraphy generation: A systematic literature review.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwen2.5-vl.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Abdelhay Zoizou, Arsalane Zarghili, and Ilham Chaker. 2022. Moj-db: A new database of arabic historical handwriting and a novel approach for subwords extraction. *Pattern Recognition Letters*, 159:54–60.

A Appendix

This appendix provides supplementary material for our study of Arabic calligraphy understanding in multimodal models. It includes: (1) an overview of related work situating DuwatBench within existing resources; (2) detailed dataset statistics; (3) verification and validation guidelines used by annotators and experts; (4) additional evaluation results; and (5) supplementary statistical analysis highlighting the dataset's stylistic and cultural diversity. Together, these materials emphasize DuwatBench's role in bridging linguistic meaning with visual heritage and providing a solid foundation for model evaluation.

B Related Work

Research on Arabic calligraphy intersects with Natural Language Processing (NLP), multimodal learning, and cultural heritage. Although prior work has produced valuable benchmarks and datasets, most focus on clean text, structured documents, or narrow stylistic slices, leaving the artistic and cultural complexity of calligraphy largely unaddressed. We group related work into three areas: (i) Arabic multimodal benchmarks, (ii) heritage and cultural resources, and (iii) Arabic calligraphy datasets.

B.1 Arabic Multimodal Benchmarks

Arabic multimodal benchmarks have evolved from general-purpose evaluation suites to more specialized resources. CAMEL-Bench established a large-scale benchmark covering eight domains with over 29K native-curated questions (Ghaboura et al., 2025a). Dialectal variation was later emphasized through resources such as Dallah (Alwajih et al., 2024) and JEEM (Kadaoui et al., 2025), underscoring the importance of capturing linguistic diversity across the Arab world. Building on these directions, the ARB benchmark (Ghaboura et al., 2025b) introduced step-by-step multimodal reasoning across 11 domains, spanning visual, OCR, and document understanding challenges. Although these resources significantly advance Arabic multimodal research, they remain centered on language, dialect, and document-level OCR. None address the stylistic and artistic dimensions of Arabic script, where the written form functions as both text and art.

B.2 Heritage and Cultural Benchmarks

Heritage-oriented benchmarks demonstrate how evaluation can extend beyond functional NLP to cultural knowledge and artistic expression. For example, Fann or Flop (Alghallabi et al., 2025) foregrounds Arabic poetry, requiring models to interpret figurative language and aesthetic conventions across eras and genres. Other initiatives, such as Sacred or Synthetic (Atif et al., 2025), PalmX (Alwajih et al., 2025b), and PEARL (Alwajih et al., 2025a), introduced shared tasks for benchmarking LLMs on Arabic and Islamic cultural knowledge, underscoring the importance of cultural grounding. These resources have advanced heritage-aware evaluation, but they largely address poetry, artifacts, or cultural reasoning, while the artistic and stylistic dimension of Arabic calligraphy remains overlooked.

B.3 Arabic Calligraphy Resources

Computational work on Arabic calligraphy has primarily targeted recognition, identification, and style classification. Early studies employed handcrafted features to distinguish between scripts (Kaoudja et al., 2020; Ezz et al., 2019), but their impact was limited by small or restricted-use datasets. Later contributions introduced larger collections to support OCR and style recognition at the word and manuscript level (Alyafeai et al., 2022; Ismail et al., 2023; Adam et al., 2017; Zoizou et al., 2022; Saeed et al., 2024), complemented by computational approaches for style representation (Kaoudja et al., 2021). Earlier computer vision methods also explored the recognition of artistic scripts (Allaf and Al-Hmouz, 2016; Salamah and King, 2018), while recent surveys highlight interest in script generation through deep learning and generative models (Sumayli and Alkaoud). More recently, multimodal efforts such as Qalam (Bhatia et al., 2024) have signaled a transition from traditional OCR to vision–language approaches for Arabic handwriting.

Despite these contributions, prior datasets remain limited: they typically frame calligraphy as OCR rather than detection, rely on cropped text instead of full artistic layouts, focus mainly on historical manuscripts, and lack semantic diversity. **DuwatBench** addresses these gaps by introducing 1,272 samples across six major styles with bounding box annotations and a wide range of textual categories. Unlike earlier corpora, its samples appear in complex artistic contexts such as inscriptions, signage, and artworks, foregrounding the visual–linguistic complexity of calligraphy in real-world settings.

C Data Statistics

To better characterize the distributional properties of our Arabic calligraphy dataset, we report descriptive statistics across both stylistic and thematic dimensions. Figures 6 and 7 illustrate that the corpus is primarily composed of widely used scripts such as Thuluth and Diwani, with comparatively lower representation of Kufic, Naskh, Ruq'ah, and Nasta'liq.

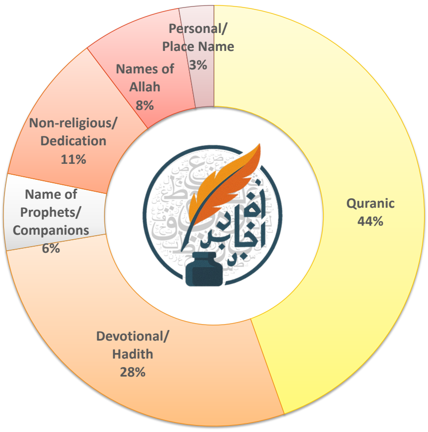

Figure 6: **DuwatBench Theme-level Distribution.** Distribution of DuwatBench textual content across thematic categories, including non-religious text, Quranic verses, devotional invocations, names of the Prophet and companions, Names of Allah, and smaller groups such as person names or places, dedications, and hadith.

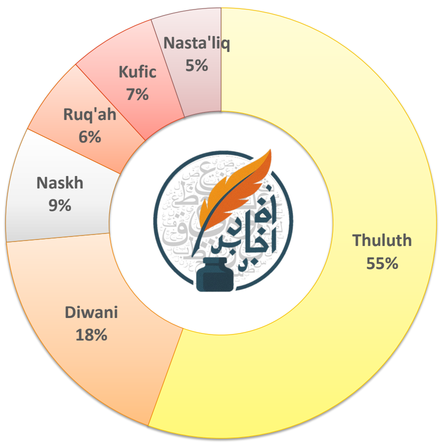

Figure 7: **DuwatBench Style-level Distribution.** Distribution of DuwatBench samples across six major Arabic calligraphy styles, with Thuluth and Diwani forming the majority and Kufic, Naskh, Ruq'ah, and Nasta'liq represented in smaller proportions.

On the content side, the dataset includes both non-religious expressions and religious material, with a strong emphasis on Quranic verses, devotional phrases, divine names, and prophetic references. This distribution underscores the dual role of Arabic calligraphy as both a cultural and sacred medium, while capturing the diversity needed for robust multimodal evaluation.

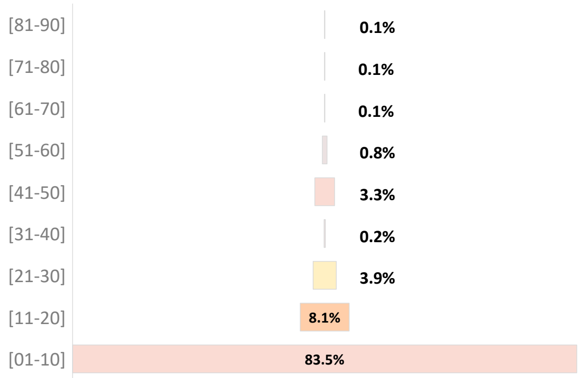

Figure 8: **Distribution of total word counts per entry in DuwatBench.** While most samples in DuwatBench consist of single words (83.5%), the dataset also includes longer inscriptions, particularly Quranic verses and combined passages.

Figure 8 illustrates the distribution of word counts across entries in DuwatBench. The majority of samples are single words, reflecting the common use of calligraphy for standalone terms such as names, motifs, or short expressions. At the same time, the dataset captures longer and more complex inscriptions, especially in religious contexts. Examples include extended Quranic verses such as Ayat al-Kursi and Al-Fatihah, as well as composite renderings that combine multiple surahs—for instance, An-Nas, Al-Falaq, and Al-Ikhlas presented alongside Ayat al-Kursi. This range highlights the dataset's ability to represent both concise and extended forms of calligraphy, ensuring coverage of diverse inscription practices.

D Verification and Annotation Guidelines

To ensure consistency and reliability across annotations, we designed a clear set of guidelines for annotators and validators. These rules covered image inclusion, box creation, transcription practices, and style/theme labeling, followed by a two-tier verification process. Table 5 summarizes the rubric applied during the annotation and verification phase of DuwatBench.

E Additional Results and Analysis

To complement the main evaluation, this section presents additional quantitative results across Arabic calligraphic styles using the chrF metric, offering a finer view of character-level recognition and how performance varies with script complexity and data representation.

Table 6 shows notable variation in recognition difficulty. Kufic records the lowest scores overall,

Aspect	Guidelines
Image inclusion	- Good resolution. - Calligraphy must be complete, legible, and undistorted. - Reject blurred, overexposed, cropped, or incomplete text. - Prefer authentic artistic backgrounds preserving natural context.
Bounding boxes	- Draw boxes tightly around each word. - Do not cut diacritics or ornamental strokes. - Avoid excessive empty space around text.
Transcription	- Manually transcribe all visible text. - Normalize letter forms and diacritics for consistency. - Use Unicode-compliant Arabic script.
Style & theme labeling	- Assign one of six calligraphic styles: Thuluth, Diwani, Kufic, Naskh, Ruq'ah, Nasta'liq. - Assign a thematic category: non-religious, Quranic, devotional, Names of Allah, prophetic references, or smaller groups (personal dedications, hadith).
Quality control (Tier 1)	- Check resolution, completeness of text, and absence of artifacts.
Expert validation (Tier 2)	- Verify transcription accuracy and diacritic placement. - Confirm style and theme classification. - Resolve disagreements through consensus.

Table 5: **DuwatBench Verification and Annotation Guidelines.** The table summarizes the procedures followed during dataset verification and labeling. It outlines key aspects of the annotation process from image inclusion criteria and bounding box placement to transcription, style, and thematic labeling, alongside a two-tier validation framework. Tier 1 ensures visual and structural quality control, while Tier 2 involves expert validation of transcription accuracy, diacritic integrity, and style–theme consistency.

	Models	Kufic	Thuluth	Diwani	Naskh	Ruq'ah	Nasta'liq
Open-source	Llava-v1.6-mistral-7b-hf (Liu et al., 2023)	7.1477	10.0517	7.1070	8.5065	7.7523	10.4573
	EasyOCR (Kittinaradorn et al., 2024)	11.0156	9.5411	5.5714	37.5340	19.9125	16.0816
	InternVL3-8B (Chen et al., 2024)	15.5629	23.4338	24.2987	21.4362	11.6528	15.9995
	Qwen2.5-VL-7B (Team, 2025)	21.5025	37.6000	30.4223	57.2947	37.8429	37.4302
	Qwen2.5-VL-72B-Instruct (Team, 2025)	24.0088	<u>45.4786</u>	35.1459	<u>65.2360</u>	<u>49.1668</u>	<u>42.4100</u>
	Gemma-3-27B-IT (Team et al., 2025)	37.1295	52.6964	45.8464	69.3757	47.0942	53.8704
	trocr-base-arabic-handwritten* (Microsoft, 2022)	1.2777	1.5063	2.5502	1.2485	3.3519	2.3474
	MBZUAI/AIN* (Heakl et al., 2025)	<u>31.7205</u>	40.7797	<u>40.0978</u>	61.6722	55.6625	42.1033
Closed-source	claude-sonnet-4.5 (Anthropic, 2025)	23.6520	48.6203	30.4111	57.0519	30.8645	40.8687
	gemini-1.5-flash (Reid et al., 2024)	<u>42.9874</u>	<u>63.3661</u>	<u>57.5302</u>	80.5358	71.9229	68.9132
	gemini-2.5-flash (AI, 2025)	47.4052	77.0677	62.1220	77.8096	<u>69.2090</u>	71.2734
	gpt-4o-mini (OpenAI, 2024)	27.9043	40.8158	35.0949	71.6135	52.1796	51.8428
	gpt-4o (OpenAI, 2024)	35.0349	55.8203	50.5606	<u>80.2030</u>	67.5542	<u>69.8079</u>

Table 6: **ChrF scores across Arabic calligraphy styles on DuwatBench.** ChrF values are reported per script style (Kufic, Thuluth, Diwani, Naskh, Ruq'ah, Nasta'liq) for open and closed-source models. Higher scores indicate better recognition quality. *Note:* * indicates Arabic-specific models. Values in blue denote the best performance, while underlined values indicate <u>the second-best</u> within each category (open or closed-source).

reflecting its geometric rigidity and sparse representation in training data. Thuluth and Diwani achieve moderate results—lower for open-source models but substantially higher for closed-source ones such as Gemini-2.5-Flash and Gemini-1.5-Flash. Naskh and Ruq'ah remain the most accurately recognized, while Nasta'liq yields moderate but inconsistent outcomes, performing better in multilingual models like Gemma-3-27B-IT, Gemini-2.5-Flash, and GPT-4o. Notably, MBZUAI/AIN demonstrates stable performance across scripts, leading in Ruq'ah and achieving competitive results in Diwani and Kufic, reflecting effective script-aware adaptation despite its comparatively smaller model size. These observations indicate that style complexity and data diversity, rather than model scale alone, play a decisive role in determining recognition quality for Arabic calligraphy.

Metric	Mean	Std	Min	Max	Range / Rel. Impr.
CER_mean	0.6456	0.1993	0.3700	0.9932	0.6232 (62.75%)
WER_mean	0.7434	0.1819	0.4478	0.9998	0.5520 (55.21%)
NLD_error	0.5940	0.2017	0.3166	0.9632	0.6466 (67.13%)
chrF	38.2873	21.3941	1.7938	71.8174	70.0236
ExactMatch	0.1791	0.1390	0.0000	0.4167	0.4167 (41.7%)
t-test (Open vs Closed)	CER_p=0.0181	WER_p=0.0087	NLD_p=0.0112	chrF_p=0.0084	ExactMatch_p=0.0029

Table 7: **Cross-Model Error Statistics and Significance Analysis.** Descriptive statistics and t-test comparisons between open and closed-source models on DuwatBench. Results show that closed-source models achieve significantly lower normalized edit distance and higher ExactMatch accuracy, indicating greater robustness to stylistic variation in Arabic calligraphy.

F Supplementary Statistical Analysis

This section provides additional quantitative analysis of model performance on DuwatBench. It complements the main results by presenting descriptive statistics, relative improvements, and statistical comparisons between open and closed-source systems. Together, these analyses provide a deeper understanding of how different models handle stylized Arabic calligraphy across recognition metrics.

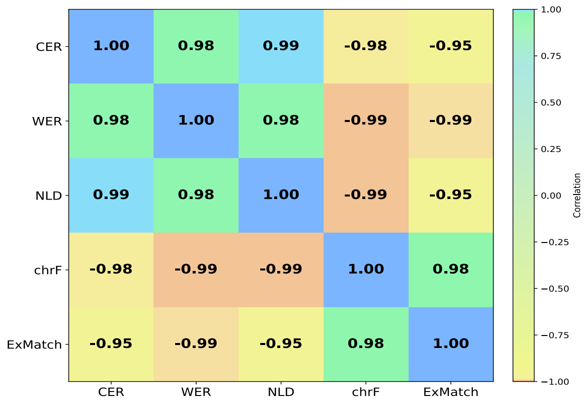

Figure 9: **Metric Correlation Heatmap.** Correlation matrix across five metrics on DuwatBench, showing that higher edit distances (CER, WER, NLD) strongly inversely correlate with chrF and ExactMatch, confirming their complementary behavior in assessing recognition quality.

Across all evaluated models, the average CER (0.65), and average WER (0.74) reflect the inherent difficulty of recognizing visually complex Arabic scripts. The mean NLD (0.59) indicates frequent partial mismatches, while mean chrF (38.29) and ExactMatch (0.18) suggest limited fluency and full-sequence accuracy. Large standard deviations ($\approx$ 0.18–0.2 for CER/WER/NLD and $\approx$ 21 for chrF) reveal substantial variability across systems, highlighting uneven robustness under diverse styles and artistic layouts.

Relative improvement (Rel. Impr.) between the strongest and weakest systems reaches 55–67% for CER, WER, and NLD, underscoring the impact of model scale and multimodal grounding. Independent two-sample t-test comparing open- and closed-source models (Table 7) show that closed-source systems achieve significantly lower error rates across CER, WER, and NLD ($p < 0.05$), as well as significantly higher chrF and ExactMatch scores ($p < 0.01$), indicating consistent advantages at both character- and sequence-level evaluation, with the strongest gains in normalized edit distance and exact transcription accuracy.

These findings align with the qualitative analyses presented in the main paper, where closed-source models demonstrated greater stability on ornate scripts (e.g., Diwani, Thuluth), whereas open-source models exhibited higher variance and frequent segmentation errors. Together, Figure 9 and Table 7 provide a comprehensive quantitative view of model behavior, reinforcing the need for style-aware and culturally grounded evaluation of Arabic calligraphy recognition.