The Unwritten Benchmark: A New Challenge for Multimodal Machine Learning in Abstract Perceptual Reasoning

Supplementary Material

A. Appendix

A.1. Dataset Creation: Detailed Methodology

The creation of the synthetic `WORD_DATASET` involved a two-stage process: (1) generating a clean list of common English words, and (2) programmatically synthesizing the multimodal samples from our live-recorded letter primitives.

A.1.1. Common Word List Generation

A high-quality, clean word list was essential for the benchmark's integrity. We started with a high-frequency list of common English words and applied a strict filtering protocol.

Source The initial seed list was derived from the `wordfreq` library's top English word lists, which are based on large text corpora. Any similar open-source frequency list can be used to reproduce this process; the critical component is the filtering pipeline.

Filtering Rules The raw list was cleaned using the following sequential rules:

1. **Length Filter:** Words were restricted to lengths between 2 and 5 characters, inclusive.
2. **Character Filter:** Words were required to contain only ASCII letters `[a-z]` after being converted to lowercase.
3. **Stoplist Removal:** We removed words containing digits, hyphens, apostrophes, or other non-alphabetic characters.
4. **Coverage Filter (Per-Style):** A word was excluded from a specific style's word list if any of its constituent letters were missing from that style's set of recorded primitives. This was crucial for ensuring that we could synthesize every word in a given list for that style (e.g., the 'Retrace' style has an incomplete alphabet list).
5. **Deduplication:** The final lists were deduplicated and stored in their canonical lowercase form.

This process yielded cleaned word lists for each handwriting style, ready for synthesis.

A.1.2. Word Synthesis Pipeline

The pipeline maps a word to its constituent letter clips and concatenates them to create the final multimodal sample.

File Naming and Mapping For each letter 'l' in a given style 's', the system maps to three source files. The file structure is as follows:

- **Standard (s=1):** `.../l-1.mp3`, `.../l-1.mp4`, `.../l-1_muted.mp4`
- **Cursive (s=2):** `.../l-2c.mp3`, `.../l-2c.mp4`, `.../l-2c_muted.mp4`
- **Retrace (s=3):** `.../l-3r.mp3`, `.../l-3r.mp4`, `.../l-3r_muted.mp4`

An availability check ensures all three modalities exist for every letter in a word before synthesis proceeds for that word. Styles are never mixed during synthesis. To reduce writer and style bias, the dataset was collected from three distinct participants, each contributing a consistent handwriting style, with each participant recording the style most natural to them.

Concatenation Mechanics We used a robust `ffmpeg` pipeline to concatenate clips while preserving media properties.

- **Video Concatenation:** For both muted and audio-visual videos, we used the `ffmpeg concat demuxer`. This method performs a container-level stitch without re-encoding, making it fast and lossless. A text file listing the absolute paths of the letter clips was generated for each word, and `ffmpeg` was run with the `-c copy` flag. Example command for creating a video for the word "arm":

  First, create arm.txt:
  – file '/path/to/a-1.mp4'
  – file '/path/to/r-1.mp4'
  – file '/path/to/m-1.mp4'
  ffmpeg -v quiet -y -f concat -safe 0 -i arm.txt -c copy arm_final.mp4

- **Audio Concatenation:** To maximize speed and avoid potential decoding errors in source clips, we used a raw bytewise concatenation as the primary method.

  Example command for creating audio for the word "arm"
  cat a-1.mp3 r-1.mp3 m-1.mp3 > arm_final.mp3

Timing No artificial spacing was introduced between letters; the clips are concatenated back-to-back. Future work could explore the impact of inserting fixed-duration silences (audio) and freeze-frames (video) to simulate natural pauses.

Quality Control To ensure the integrity of the synthesized dataset, every letter primitive was manually verified for synchronization during dataset construction. In addition, random synthesized word samples were manually spot-checked after concatenation to confirm that the stitching process did not introduce noticeable artifacts, corruption, or modality misalignment. These checks were intended to ensure that model behavior reflected the difficulty of the task itself rather than errors in preprocessing or file construction.

A.2. Human Evaluation: Additional Details

The human evaluation was designed to establish a robust performance ceiling and gather qualitative insights into how humans solve this task.

Participant Testimonies and Observations Beyond the quantitative scores, we collected qualitative feedback and made observations that informed our analysis:

- **Embodied Simulation:** We frequently observed participants instinctively mimicking the writing motions with their own fingers while watching the video clips. This suggests that human perception on this task is not merely passive observation but an active, *embodied simulation* used to better understand and internalize the kinematic patterns. This behavior also appeared in audio-only trials, where participants reported relying on the rhythm and structure of the pen sounds to mentally reconstruct plausible writing motions.
- **Audio-Only Inference Cues:** Human performance in the audio-only condition, while much lower than video-based performance, was still above trivial guessing. This appears to be supported by genuine acoustic structure in the signal, including cues such as stroke count, pauses, timing, and intensity changes that correspond to properties of specific letters or writing patterns. Participant feedback suggested that these rhythmic and temporal cues sometimes enabled partial inference even in the absence of visible motion.
- **Familiarity Bias:** Several participants noted that it was significantly easier to guess words written in a style that matched their own (Standard vs. Cursive). This points to a reliance on familiar, internalized motor programs for inference, and likely helps explain the lower human performance on the British Cursive condition relative to Standard and Retrace.
- **Contextual Reinforcement:** One participant provided a particularly insightful comment on the 'Retrace' style: *"it was easiest to guess the retraced words because the sample would show the person going over strokes...which gave us more context and time to understand what was being written."* This confirms our hypothesis that humans leverage repetition for high-level contextual clarity.

A.3. Prompting Strategy

Initial exploratory experiments revealed that open-ended prompts resulted in models frequently defaulting to common, short words (e.g., "the," "cat"), a form of mode collapse. To elicit more meaningful responses and conduct a more rigorous evaluation, we incorporated a length constraint into the prompt. After significant prompt engineering, the following formats were finalized for all evaluations. The variable n was programmatically populated with the correct word length for each sample.

The numeric constraint was used only to restrict the expected output length, not to provide semantic information about the answer itself. In other words, the prompt does not narrow the content of the response beyond the number of letters; it simply prevents degenerate outputs and makes evaluation more comparable across samples.

Final Prompts

Style note appended to all prompts:
The handwriting style may be American standard print, British cursive, or retrace (letters may be traced over). Only one of the styles is used in this sample.

Prompt for **Audio** Files:
Listen to the pen-on-paper audio of someone writing. The word has n letters. What English word is being written? Answer with ONE lowercase word (a-z). If unsure, guess the word. Do not apologize. Do not explain. Return only the word." + Style note

Prompt for **Video** Files:
Watch the handwriting. The word has n letters. What English word is being written? Answer with ONE lowercase word (a-z). If unsure, guess the word. Do not apologize. Do not explain. Return only the word." + Style note

The above prompts provided to the MLLMs were finalized after a thorough iterative process designed to minimize ambiguity and prevent common failure modes like mode collapse. Some of the prompts tried and later discarded are as follows:

V1: Initial Open-EndedPrompt - Discarded

"What word is being written in this video?"

Result: This prompt consistently led to total failure as all models started apologizing and no meaningful results were produced in preliminary tests.

V2: Initial Open-Ended Prompt with Answer length limit - Discarded

> "What word is being written in this video? Answer with ONE lowercase word (a-z). Do not apologize."

Result: This prompt consistently led to models defaulting to short, high-frequency words ("the," "a," "is"), regardless of the input's duration or complexity.

V3: Added Style Note - Discarded

> "Analyze the handwriting in the video. The style could be standard, cursive, or retrace. What word is being written?"

Result: This prompt still suffered from significant mode collapse. The style note alone was insufficient to guide the models toward a more detailed analysis.

V4: Prompt with Length Constraint - Discarded

> Watch the handwriting. The word has n letters. What English word is being written? Answer with ONE lowercase word (a-z).

Result: The inclusion of the word length {n} was the most critical factor in eliciting meaningful responses, forcing the models to move beyond their default priors and attempt a genuine interpretation of the input signal.

Finally, a combination of V3 and V4 was tried, which yielded the best results so far, and was thus used to form the Final prompts

A.4. Handwriting Style Guidance

To ensure consistency in the foundational letter primitives, participants were guided by standard handwriting worksheets commonly used in elementary education. For the 'American Standard' and 'British Cursive' styles, participants referred to worksheets that provided stroke-order diagrams and dotted outlines for each uppercase and lowercase letter.

Figure 1. Stroke pattern for American Standard English that was followed during data collection.

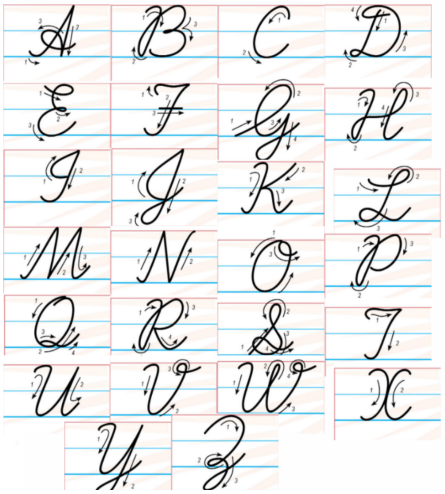

Figure 2. Stroke pattern for British Capital cursive that was followed during data collection.

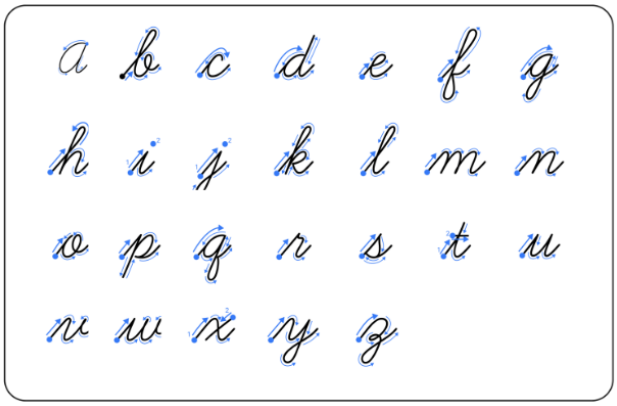

Figure 3. Stroke pattern for British Small cursive that was followed during data collection.

A.5. Ethics and Reproducibility Statements

Ethics Statement: All data was collected from consenting adult participants who were informed of the study's purpose. The dataset was fully anonymized, with no personally identifiable information collected or stored. The goal of this research is to advance scientific understanding of AI capabilities and limitations. We do not foresee any direct negative societal impacts, and the dataset does not contain any sensitive content.

Reproducibility Statement: The Unwritten Benchmark dataset has been made publicly available on our Project Page. All models were accessed via their APIs or HuggingFace, and the exact model versions are noted in the main text. The prompts used for evaluation are provided in their entirety in the paper to ensure that our results are fully reproducible.