Learning to Align: Addressing Character Frequency Distribution Shifts in Handwritten Text Recognition

Panagiotis Kaliosis[1,2,†] **John Pavlopoulos**[2,3]

[1]Stony Brook University, New York, USA
[2]Archimedes, Athena Research Center, Greece
[3]Athens University of Economics and Business, Greece
pkaliosis@cs.stonybrook.edu, annis@aueb.gr

Abstract

Handwritten text recognition aims to convert visual input into machine-readable text, and it remains challenging due to the evolving and context-dependent nature of handwriting. Character sets change over time, and character frequency distributions shift across historical periods or regions, often causing models trained on broad, heterogeneous corpora to underperform on specific subsets. To tackle this, we propose a novel loss function that incorporates the Wasserstein distance between the character frequency distribution of the predicted text and a target distribution empirically derived from training data. By penalizing divergence from expected distributions, our approach enhances both accuracy and robustness under temporal and contextual intra-dataset shifts. Furthermore, we demonstrate that character distribution alignment can also improve existing models at inference time without requiring retraining by integrating it as a scoring function in a guided decoding scheme. Experimental results across multiple datasets and architectures confirm the effectiveness of our method in boosting generalization and performance. We open source our code at this link.

1 Introduction

Handwritten Text Recognition (HTR) enables the automatic conversion of handwritten input into machine-readable text, supporting critical applications such as historical manuscript digitization, document analysis, and archival research. Although often overlooked, systematic shifts in character frequency distributions, caused by linguistic evolution, orthographic reforms, and contextual usage, are an important consideration when building robust HTR systems across diverse time periods or regions (Moreno, 2005). In the transition from Old English to Modern English, for example, grammatical changes, such as the loss of case markings, led to notable alterations in letter distributions (Moreno, 2005). Character frequency distributions also differ across languages that share the same alphabet, with measurable disparities even among closely related languages (Grigas and Juskeviciene, 2018). Beyond language, writing style and genre exert substantial influence on letter usage, which is reflected in systematic differences across news articles, novels, and scientific texts (Zhao and Zheng, 2024). Specialized domains further introduce deviations: for instance, proprietary prescription drug names diverge markedly from standard English distributions (Carico et al., 2022), while authorship attribution studies exploit subtle statistical variations in letter frequencies to differentiate between writers (Diurdeva et al., 2016; Merriam, 1994; Kjell, 1995). These findings underscore the dynamic nature of character frequency distributions and their dependence on both linguistic and contextual factors.

HTR models face challenges when transcribing historical manuscripts, where orthographic conventions and character usage shift considerably over time (Tsochatzidis et al., 2021; Cascianelli et al., 2022). A common mitigation strategy involves training on diverse corpora to increase generalization, but this does not necessarily yield optimal performance (Nguyen et al., 2022). In initial experiments, we observed that models trained on heterogeneous data tend to underperform in specific subsets where character frequency distributions deviate, leading to systematic biases and suboptimal transcription accuracy, a phenomenon that to the best of knowledge has not been thoroughly explored in prior work.

To address these challenges, we propose FADA (Frequency-Aware Distribution Alignment), a novel training framework that aligns a model's predicted character-level frequency distributions with empirically observed ones. When such distributions are available, FADA enhances robustness to

[†]Work done during internship at Archimedes Research Center before joining Stony Brook University.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18667–18684
November 4-9, 2025

temporal and contextual variation by incorporating a distance-based loss function that penalizes discrepancies between predicted and empirical character frequencies during training. This alignment loss allows the model to learn general task-relevant representations while also reflecting the expected subset-specific statistical patterns. In addition, we introduce a guided decoding algorithm that integrates the frequency-aware objective into the beam search process at inference time. By treating empirical character frequency distributions as soft constraints, our method refines predictions to better match expected patterns, improving transcription accuracy without the need for retraining. FADA is the first framework to explicitly address character-level frequency distribution shifts in text recognition. In contrast to prior work, which largely overlooks intra-dataset variation, FADA introduces both a trainable alignment mechanism and an inference-time correction strategy, providing a comprehensive solution for mitigating distributional shifts.

Our main contributions are: (i) We introduce FADA, a novel training framework that explicitly addresses character-level frequency distribution shifts in recognition tasks. (ii) We propose a guided decoding strategy that incorporates empirical frequency priors into beam search, enhancing alignment between predicted and expected character frequency distributions at inference time. (iii) We demonstrate the effectiveness and generality of FADA through extensive experiments on HTR, showing consistent performance gains across multiple datasets, languages, and model architectures.

2 Related Work

Feature and distribution alignment have been extensively explored as strategies to mitigate performance degradation caused by distribution shifts in recognition models. Prior work addresses these challenges by enhancing alignment at various stages of the recognition pipeline, including structural feature extraction, semantic representation learning, and distribution matching.

In **scene text recognition (STR)**, Hu et al. (2024) proposed shape-driven attention to guide models toward higher-quality character features using geometric priors. In **vision-language models (VLMs)**, distributional alignment has also been explored (Cho et al., 2023). For example, Li et al. (2025) minimized distributional gaps between positive and negative captions to improve image-text alignment by promoting stronger visual grounding.

In **speech-related tasks**, alignment techniques have been applied to adapt to domain-specific variation. Zhou et al. (2024) investigated distribution alignment for multi-genre speaker recognition, showing that Within-Between Distribution Alignment (WBDA) (Hu et al., 2022) improved robustness but did not fully eliminate genre-specific variability. Moreover, Hou et al. (2021) aligned the character-level audio representations between a source and a target domain in an unsupervised manner via Maximum Mean Discrepancy.

Standard decoding strategies like greedy or beam search often lack mechanisms to enforce **external constraints**, which are crucial in tasks requiring adherence to linguistic structure or statistical properties. To address this, **guided decoding** methods steer generation based on predefined constraints (Wang and Shu, 2025), typically grouped as semantic, structural, or lexical (Zhang et al., 2023). Semantic constraints enforce stylistic or topical alignment (Yang et al., 2018; Ghazvininejad et al., 2017; Pascual et al., 2021), or integrate task-specific priors (Kaliosis et al., 2024; Samprovalaki et al., 2024; Chatzipapadopoulou et al., 2025). Structural approaches promote syntactic coherence (Bastan et al., 2023; Lu et al., 2021), while lexical constraints guide the inclusion or exclusion of specific tokens (Anderson et al., 2017; Hokamp and Liu, 2017; Yao et al., 2024).

Despite this progress, existing approaches do not explicitly address the impact of character-level frequency distribution shifts, which are particularly prominent in recognition tasks involving historical or domain-specific data. Most alignment methods focus on feature-level or semantic alignment, overlooking statistical priors at the character level that can guide model predictions. **FADA** fills this gap by introducing a frequency-aware training and decoding framework that directly incorporates character-level distributional knowledge into both learning and inference.

3 The Proposed FADA Method

Our proposed method, FADA, introduces a novel training framework that aligns the predicted character-level frequency distribution of the model with an empirical target distribution at each training step, thereby reducing discrepancies caused by temporal and contextual shifts.

For example, if the model predicts a character

frequency of 9.3% for the letter "a", while the empirical distribution indicates an expected 11.1%, FADA encourages the model to adjust its predictions to reduce this gap. To achieve this, we introduce an auxiliary loss term that penalizes divergence between the predicted and empirical character frequency distributions. This alignment loss is jointly optimized with the task-specific objective (e.g., CTC or Cross-Entropy loss), enabling the model to preserve recognition performance while adapting to subset-specific distributional properties. As a result, the model becomes more robust to intra-dataset shifts and better aligned with the statistical patterns present in the data.

Empirical Relative Frequency Distributions To capture character frequency variations, we compute empirical relative frequency distributions from the training data. First, we normalize transcriptions by converting all text to lowercase to ensure consistency. Frequencies are calculated by counting character occurrences and normalizing by the total number of characters. Intra-dataset distribution shifts are systematic variations in character frequency distributions across distinct subsets of a corpus. To model such shifts, we compute separate frequency distributions for each temporal or contextual subset of the training set (e.g., distinct centuries/regions).

Character Frequency Analysis Figure 1 illustrates the character-level relative frequency distributions of the HPGT dataset (Platanou et al., 2022) over seven centuries, revealing clear distributional shifts in character usage. Certain letters, such as δ, ϰ, and γ, display stable frequencies over time, suggesting consistent orthographic roles in written Greek. In contrast, others, such as τ and ε, exhibit notable fluctuations, suggesting potential shifts in orthographic conventions or variations in dataset composition. Additionally, characters such as ϖ̃ and ἔ, are absent in earlier centuries but emerge in later ones, which may reflect historical linguistic developments or artifacts of dataset annotation and coverage. However, these variations are not solely attributable to linguistic evolution; external factors, such as dataset biases or uneven temporal representation, may also influence the observed distributions. We present a detailed analysis including correlation scores between the character frequency distributions across centuries in Appendix A.

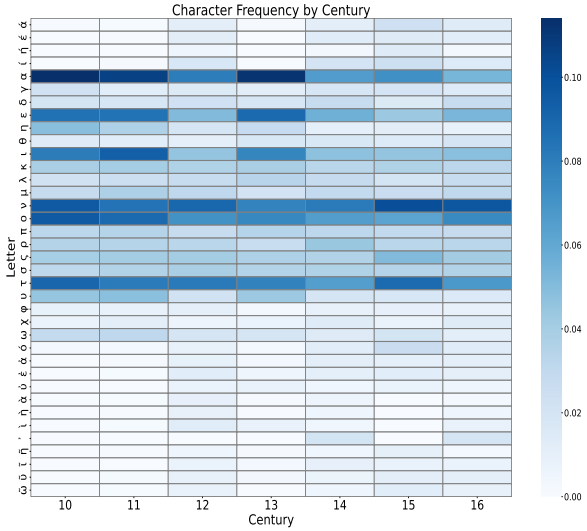

Figure 1: Empirical character-level frequency distributions in the HPGT dataset across seven centuries (10th–16th CE). Characters with relative frequency below 0.01 are omitted for clarity.

FADA Training Framework (FADA$_{TR}$) In conventional recognition models, training is typically guided by task-specific objectives such as CTC or CE loss. While these objectives effectively optimize transcription accuracy, they do not account for intra-dataset shifts in character frequency distributions. As a result, models trained on aggregated datasets may develop implicitly biases toward dominant character distributions, leading to suboptimal performance when applied to subsets that exhibit different statistical properties at the character level.

To address this limitation, we introduce an auxiliary *distribution alignment loss* that encourages the model's predicted character frequency distributions to align with empirical distributions observed in the training data. At each training step, we compute the predicted distribution from the model's output logits and compare it with a sample-specific empirical distribution derived from the training set (e.g., based on the sample's century or region). The alignment loss penalizes discrepancies between the predicted and empirical distributions, guiding the model toward representations that preserve the expected statistical properties of the corresponding subset. This loss is designed to complement—not replace—the primary recognition objective (CTC or CE), thereby enhancing the model's robustness to intra-dataset distributional shifts without compromising transcription accuracy.

To formally define this alignment loss, we employ the *Wasserstein distance*, a well-established

metric for quantifying discrepancies between probability distributions. Given two probability distributions D and Q , the Wasserstein-p (W_p) distance is defined as $W_p(D, Q) = \left(\frac{1}{n}\sum_{i=1}^{n} \|D_{(i)} - Q_{(i)}\|^p\right)^{\frac{1}{p}}$ where $D_{(i)}$ and $Q_{(i)}$ denote the i-th elements of the sorted (in ascending order) versions of D and Q, respectively, and n is the number of characters. Since we operate over discrete relative frequency distributions, i.e., one-dimensional empirical probability vectors summing to one, we specifically adopt the second-order Wasserstein distance (W_2), where $p = 2$:

$$\mathcal{L}_{W_2} = W_2(D, Q) = \left(\frac{1}{n}\sum_{i=1}^{n} \|D_{(i)} - Q_{(i)}\|^2\right)^{\frac{1}{2}} \tag{1}$$

This choice is motivated by the fact that W_2 corresponds to the Euclidean norm, offering a natural and computationally efficient measure of distributional divergence that is sensitive to fine-grained differences. In the one-dimensional case, W_2 simplifies to the Root Mean Squared Error (RMSE) between the sorted frequency vectors, making it especially suitable for character-level distributions. At each training step, the total loss is defined as a convex combination of the primary recognition loss and the distribution alignment loss. The former corresponds to either CTC ($\mathcal{L}_{\text{CTC}}$) or Cross-Entropy ($\mathcal{L}_{\text{CE}}$) loss, depending on the model architecture. The latter is the alignment loss $\mathcal{L}_{W_2}$ introduced in Equation 1.

FADA Guided Decoding Framework (FADA$_{\text{GD}}$) Our proposed method also supports an inference-time application via a guided decoding algorithm that aligns predicted character-level frequency distributions with empirical targets. While training-time alignment enhances robustness to distributional shifts, inference-time adjustment provides an additional mechanism for refining predictions *without* requiring model retraining. To enable this, we incorporate a frequency-aware scoring component into the beam search process, which biases candidate sequences toward those whose character distributions more closely match the expected empirical ones. In this way, we correct potential distributional mismatches at inference time, thereby improving overall transcription accuracy.

In standard beam search, candidate sequences are ranked based on their cumulative log-probability, favoring highly probable continuations. We extend this decoding objective by introducing a penalty term that quantifies the divergence between a candidate sequence's predicted character frequency distribution and the corresponding empirical distribution. This penalty term is applied at each decoding step, enabling the beam search process to dynamically incorporate character-level alignment as the sequence is generated. Specifically, at decoding step t, each candidate sequence $V = (v_1, v_2, \ldots, v_t)$ is scored using the following frequency-aware objective:

$$\mathcal{S}_{\text{GD}} = [\lambda \cdot \sum_{i=1}^{t} \log P(v_i|v_{<i}) - (1-\lambda) \cdot W_2(D_{\text{pred}}, D_{\text{emp}})] \tag{2}$$

Here, $P(v_i|v_{<i})$ denotes the model-assigned probability of character v_i given the preceding sequence $v_{<i}$. D_{pred} and D_{emp} represent the predicted and empirical character frequency distributions, respectively. The first term captures the standard beam search objective and the second imposes a penalty based on the Wasserstein distance W_2 (see Equation 1). The hyperparameter $\lambda \in [0, 1]$ controls the trade-off between maximizing likelihood and enforcing distributional alignment.

This formulation ensures that candidate sequences with both high likelihood and strong agreement with the expected character frequency distribution are favored during decoding. Importantly, at each decoding step, the alignment penalty is computed over the entire (potentially incomplete) candidate sequence, rather than applied at the individual token level. While a per-token penalty may seem intuitive, it treats frequency alignment as a local property and risks biasing the model toward optimizing each next-token prediction in isolation. In contrast, our approach promotes global coherence by encouraging sequences whose overall character-level statistics align better with the empirical target distribution. As a result, the guided decoding strategy functions as a lightweight calibration mechanism that leverages known statistical properties of the target domain to enhance recognition performance without additional training. Algorithm 1 presents both our training-time and our inference-time alignment framework in pseudo-code form.

4 Experiments

We evaluate the performance of FADA primarily on HTR, using two model architectures across three

Algorithm 1 The proposed FADA framework

▷ Train set w/ domain labels d_i (e.g., century): $\mathcal{D} = \{(x_i, y_i, d_i)\}$

▷ Target character distribution per domain d: $\{p^{(d)}_{\text{target}}\}$

▷ Model with parameters θ: f_θ

▷ Loss weighting factor: λ

▷ Character Distribution calculation function: CharDist

Frequency-Aware Training (FADA$_{\text{TR}}$)

for each minibatch $\mathcal{B} = \{(x_i, y_i, d_i)\} \subset \mathcal{D}$ **do**
  $y_{\text{pred}} \leftarrow f_\theta(x)$ ▷ Model predictions
  $\mathcal{L}_{\text{CE}} \leftarrow \text{CE}(y_{\text{pred}}, y)$ ▷ Cross-entropy loss
  ▷ Compute alignment loss across samples
  $\mathcal{L}_{\text{FADA}} \leftarrow 0$
  for each (y_i^{pred}, d_i) in (y_{pred}, d) **do**
    $q_i^{\text{pred}} \leftarrow \text{CharDist}(y_i^{\text{pred}})$
    $\mathcal{L}_{\text{FADA}} \mathrel{+}= W_2(q_i^{\text{pred}}, p^{(d_i)}_{\text{target}})$
  end for
  $\mathcal{L}_{\text{FADA}} \leftarrow \mathcal{L}_{\text{FADA}}/|\mathcal{B}|$
  ▷ Combine losses
  $\mathcal{L}_{\text{total}} \leftarrow \lambda \cdot \mathcal{L}_{\text{CE}} + (1-\lambda) \cdot \mathcal{L}_{\text{FADA}}$
  $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}$ ▷ Gradient update
end for

Frequency-Aware Guided Decoding (FADA$_{\text{GD}}$)

for each beam search decoding step **do**
  beamScores = dict()
  for each beam search sequence h **do**
    $d \leftarrow d_i$ ▷ e.g., target century/region
    ▷ Get LM decoder's score for h
    $\mathcal{S}_{\text{D}}(h) = \sum_{i=1}^{|h|} \log P(h_i | h_{<i})$
    ▷ Compute alignment score (Eq. 2)
    $q_h \leftarrow \text{CharDist}(h)$
    $\mathcal{S}_{\text{align}}(h) \leftarrow W_2(q_h, p^{(d)}_{\text{target}})$
    ▷ Compute FADA$_{\text{GD}}$ score
    $\mathcal{S}_{\text{GD}}(h) \leftarrow \lambda \cdot \mathcal{S}_{\text{D}}(h) - (1-\lambda) \cdot \mathcal{S}_{\text{align}}(h)$
    ▷ Update beam score
    beamScores$[h] \leftarrow \text{S}_{\text{GD}}(h)$
  end for
end for

datasets, two real-world and one synthetic, spanning Greek and French manuscripts from the 10th to 16th cent. CE. To demonstrate the broader applicability of our method, we also conduct secondary experiments on Automatic Speech Recognition (ASR) using the Whisper model (Radford et al., 2023) and an English dataset. Next, we discuss dataset and model details (§4.1-§4.2), baselines (§4.3), evaluation metrics (§4.4), and then we present the experimental results (§4.5), followed by an ablation study and qualitative analyses.

4.1 Datasets

HPGT Dataset (HTR): HPGT (Platanou et al., 2022) is a Handwritten Paleographic Greek Text recognition dataset comprising 77 digitized page-level images from Greek manuscripts in the Oxford University Bodleian Library.[1] The images are segmented into 1,698 line-level between the 10th to 16th century CE. See Appendix A for more details about the intra-dataset character distributions and correlations.

CATMuS-FR Dataset (HTR): CATMuS (Clérice et al., 2024) is a multilingual manuscript dataset containing over 200 historical texts. We extracted all French-written samples to form CATMuS-FR, consisting of 3,401 samples spanning four centuries (13th to 16th) and six distinct script types. See Appendix B for more details about the intra-dataset character distributions and correlations.

Synthetic Dataset (HTR): To evaluate FADA under controlled conditions, we generated a synthetic Greek dataset with deliberately induced shifts in character frequency distributions. Using predefined prompts and topics, we synthesized text with four different LLMs (GPT-4 (OpenAI, 2023), Llama-3.1 (Meta-AI, 2024), Gemini (Google, 2024), and Claude-3.5 (Anthropic, 2024)) each producing outputs with distinct statistical properties, effectively simulating four separate "centuries." A total of 1,690 text segments were carefully curated and rendered as images using an open-source handwritten text generator.[2] More details about the dataset creation are provided in Appendix C.

EdAcc Dataset (ASR): EdAcc (Sanabria et al., 2023) is an ASR dataset containing 40 hours of transcribed speech from speakers with diverse linguistic backgrounds and English accents. We selected the four most represented categories; US English, Southern British English, Irish English, and Indian

[1] https://bav.bodleian.ox.ac.uk/greek-manuscripts

[2] https://github.com/Belval/TextRecognitionDataGenerator

English, resulting in 2,756 samples. More details about the inter-group character distribution shifts and correlation scores are provided in Appendix D.

4.2 Backbone Models

We experiment with three backbone models featuring distinct architectures across two recognition tasks. In our experiments, C-RNN is trained from scratch, while TrOCR and Whisper are fine-tuned from publicly available pretrained checkpoints. We evaluate each model under two baseline configurations (see §4.3) as well as our proposed distributional alignment framework (FADA).

C-RNN (HTR): C-RNN (Shi et al., 2015) is a widely used architecture for HTR that combines a CNN for visual feature extraction with an RNN for sequential text generation. The model outputs text at character level and is trained using the CTC loss.

TrOCR (HTR): TrOCR (Li et al., 2021) is a state-of-the-art Transformer-based (Vaswani et al., 2017) approach for HTR. It comprises a Vision Transformer (ViT) (Dosovitskiy et al., 2021) as an image encoder and a text Transformer (Vaswani et al., 2017) as the decoder. TrOCR operates at token level and is trained using CE loss.

Whisper (ASR): Whisper (Radford et al., 2023) is a state-of-the-art speech recognition model based on a Transformer-based encoder-decoder architecture. It is pre-trained on a range of speech processing tasks, including ASR and speech translation.

4.3 Baseline Configurations

Fine-tuned Backbone Model: This configuration serves as our simplest baseline. For Transformer-based models (TrOCR and Whisper), we fine-tune publicly available pre-trained checkpoints on each target dataset using only the standard task-specific objective (CE loss). For the C-RNN model, which does not rely on a pre-trained backbone, we train the model from scratch using CTC loss. This setup reflects the default recognition setting without any frequency-level supervision, and enables direct assessment of the added value introduced by our proposed distributional alignment mechanism.

Temporal/Regional Token Tagging (TR-TT): Following strategies employed in multilingual and multi-task models such as mBART (Liu et al., 2020) and Whisper (Radford et al., 2023), we fine-tune each backbone model by prepending a special token to each input sequence indicating the temporal or regional context. This token is treated as part of the input and provides an implicit conditioning signal for generation. This configuration does not incorporate any explicit alignment with empirical distributions. Instead, it relies on the model's ability to learn contextual associations from the data. Training is performed end-to-end using only the standard task-specific objective (CTC or CE).

4.4 Evaluation Metrics

Character Error Rate (CER) measures transcription accuracy at character level. It is computed as the edit distance between the predicted and the reference text, normalized by the length of the reference. Lower CER indicates better performance.

Word Error Rate (WER) is the word-level counterpart of CER. It calculates the edit distance between the predicted and reference word sequences, normalized by the number of words in the reference. As with CER, lower WER values correspond to higher transcription accuracy.

4.5 Experimental Results

We evaluate our proposed method (FADA) across multiple backbone models (§4.2) and compare it to two baseline configurations, each involving fine-tuning (or training from scratch depending on the architecture) the backbone models on the target dataset (§4.3). Our main configuration, dubbed $\text{FADA}_{\text{TR-GD}}$, fine-tunes (or trains) each model using a combined objective that includes the standard task-specific loss (CTC or CE) and our distributional alignment loss. At inference time, we apply our guided decoding algorithm to further refine outputs. This setup integrates frequency alignment at both training and inference stages, resulting in *end-to-end frequency-aware* recognition. We run each experiment three times with different seeds and report the mean and standard error of the mean (SEM) for each metric (see §5 for reproducibility).

Quantitative Analysis Table 1 reports the performance of the two models on the three HTR datasets. For each model (TrOCR, C-RNN), we report three scores per evaluation metric (CER, WER); one for each baseline approach, and one for our proposed method ($\text{FADA}_{\text{TR-GD}}$). For the latter, scores obtained with the best hyperparameter λ are reported. In Table 2, we examine the effectiveness of our proposed method in ASR. We report the performance of Whisper on the Edacc dataset across the two baseline methods and $\text{FADA}_{\text{TR-GD}}$.

The results in both tables show that $\text{FADA}_{\text{TR-GD}}$

Dataset	Model	Standard FT		FT w/ TR-TT		FT w/ FADA$_{\text{TR-GD}}$ (Ours)	
		CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
HPGT	TrOCR	26.92 (0.96)	73.82 (1.31)	26.89 (0.61)	72.53 (0.81)	**25.06** (0.91)	<u>70.14</u> (0.48)
	C-RNN	24.72 (0.38)	77.08 (0.85)	24.41 (0.55)	76.66 (0.81)	**23.89** (0.34)	<u>76.31</u> (0.77)
CATMuS	TrOCR	**9.56** (0.22)	37.02 (0.28)	9.68 (0.17)	<u>37.01</u> (0.69)	9.63 (0.25)	37.05 (0.47)
	C-RNN	17.08 (0.29)	54.79 (0.32)	27.12 (1.13)	67.61 (1.98)	**16.79** (0.52)	<u>53.65</u> (1.02)
Synthetic	TrOCR	5.94 (0.09)	17.61 (0.31)	5.48 (0.17)	17.12 (1.96)	**5.19** (0.27)	<u>16.60</u> (0.45)
	C-RNN	9.71 (0.73)	31.58 (2.73)	9.82 (0.34)	34.58 (1.78)	**8.31** (0.12)	<u>27.08</u> (0.36)

Table 1: **Main evaluation results.** We report average CER and WER (across three runs) and the SEM (in parentheses next to each score) across all datasets and models. FT indicates fine-tuning the model. Our proposed method outperforms all baselines in most cases. FADA$_{\text{TR-GD}}$ denotes our proposed approach combining training-time alignment and inference-time guided decoding. TR-TT is a token-based contextual conditioning baseline described in §4.3. The best CER and WER scores per model and dataset are in bold and underlined respectively.

Model	Standard FT	w/ TR-TT	w/ FADA$_{\text{TR-GD}}$
Whisper	27.11 / 40.02	28.88 / 44.09	**26.70** / <u>38.02</u>

Table 2: **ASR results on EdAcc.** CER / WER for Whisper-small across three configurations. Best scores per metric are bolded and underlined respectively.

consistently improves both text and speech recognition performance across a range of datasets and model architectures, demonstrating the effectiveness of character-level frequency distribution alignment in recognition tasks. Despite the single case where FADA$_{\text{TR-GD}}$ does not improve over TrOCR baseline (falling within the SEM) it generally provides substantial gains (up to three WER units), highlighting its robustness and broad applicability. These findings confirm that aligning model predictions with empirical character statistics can enhance recognition accuracy, even under diverse temporal or regional shifts.

Ablation study In addition to our full end-to-end configuration (FADA$_{\text{TR-GD}}$), incorporating alignment at both training and inference, we assess two partial variants: training-only alignment (FADA$_{\text{TR}}$) and inference-only alignment via guided decoding (FADA$_{\text{GD}}$). As can be seen in Table 4, inference-time alignment improves recognition performance in most cases, despite operating solely as a post-hoc adjustment to model outputs (see also §5). Training-time alignment yields larger improvements overall, as it directly shapes the model's internal representations by reducing character frequency discrepancies during optimization. While the combined approach (FADA$_{\text{TR-GD}}$) consistently achieves the best overall performance, training-only and inference-only alignment offer meaningful improvements at a lower computational cost.

Qualitative Analysis Table 3 presents two transcription examples, one from each task, to illustrate the qualitative impact of our proposed framework. We compare outputs from the fine-tuned backbone model (TrOCR or Whisper) with those produced by the same model enhanced with inference-time alignment (FADA$_{\text{GD}}$), training-time alignment (FADA$_{\text{TR}}$), and the full configuration combining both (FADA$_{\text{TR-GD}}$). In both examples, the full configuration (FADA$_{\text{TR-GD}}$) produced the most accurate transcriptions, closely matching the ground truth and outperforming the standard fine-tuned model as well as the individual FADA variants.

In the first case (HPGTR dataset, HTR), training-time alignment (FADA$_{\text{TR}}$) alone achieved the same CER and WER as the full configuration, suggesting that it can often be sufficient. In contrast, the second example (EdAcc dataset, ASR) illustrates a complementary effect; while neither training-only nor inference-only improved upon the baseline in isolation, their combination resulted in a significantly better transcription. This result highlights the synergistic benefit of combining improved representations (from training-time alignment) with frequency-aware decoding (via inference-time alignment). Note that while FADA is not designed to explicitly optimize for WER, improvements in CER which is our primary objective, naturally translate to better WER in most cases. In our examples, this correlation is clearly reflected, further reinforcing the robustness of our alignment-based framework.

Per-Century/Region Results To further assess the effectiveness of FADA, we evaluate model performance in terms of CER across individual centuries (for HTR) and linguistic regions (for ASR). Rather

Method	Transcription	CER	WER
Baseline (TrOCR)	φωνης εἰσπνατιον τὴν δὲ κας ην συμφόνας ἀφηκαν τήν	29.16	1.14
FADA$_{GD}$	φωνης εἰσπνατιον τὴν δὲ κας ην συμφόνας ἀφηκαν τήν	29.16	1.14
FADA$_{TR}$	φωνης εἰσπναττονται δίκας ἡν συμφώνως ἀφη καντὴν	**14.58**	**0.71**
FADA$_{TR\text{-}GD}$	φωνης εἰσπναττονται δίκας ἡν συμφώνως ἀφη καντὴν	**14.58**	**0.71**
Ground Truth	φωνῆς εἰσπράττονται δίκας ἣν συμφώνως ἀφῆκαν τήν		
Baseline (Whisper)	yeah and all i bet and like in the apartment that we have	18.18	0.23
FADA$_{GD}$	and uh i bet in like in the apartment that we have	19.98	0.30
FADA$_{TR}$	and uh i bet in like in the apartment that we have	19.98	0.30
FADA$_{TR\text{-}GD}$	yeah i know i bet in like in the apartment that we have	**3.63**	**0.07**
Ground Truth	yeah i know i bet uh like in the apartment that we have		

Table 3: Qualitative analysis of FADA on HTR and ASR tasks. The upper section presents a Greek manuscript transcription example, while the lower section showcases an English ASR transcript. Correctly restored characters are highlighted in green. CER and WER values are reported for each setting, demonstrating the improvements.

Dataset	Model	FADA$_{GD}$	FADA$_{TR}$	FADA$_{TR\text{-}GD}$
HPGT	TrOCR	−1.4%	−6.5%	**−6.9%**
	C-RNN	−0.8%	−3.2%	**−3.3%**
CATMuS	TrOCR	+1.7%	+1.6%	**+0.7%**
	C-RNN	−0.4%	−1.1%	**−1.7%**
Synthetic	TrOCR	−0.2%	−10.6%	**−12.6%**
	C-RNN	−0.1%	−12.9%	**−14.4%**

Table 4: **Ablation Study: The relative improvement in CER over standard fine-tuning across FADA configurations.** Negative values indicate error reduction. **Bold** marks the best per model/dataset (lower is better).

than reporting only aggregated scores, this analysis focuses on how well each model performs on specific temporal and regional subsets. The results reveal substantial variation in difficulty across these subsets. Notably, FADA-enhanced models (FADA$_{TR}$, FADA$_{GD}$, and FADA$_{TR\text{-}GD}$) exhibit improved and more consistent performance across different subsets compared to standard fine-tuning, effectively narrowing the gap between easier and more challenging cases. A detailed breakdown of per-century and per-region results is provided in Appendix F.

Lipogram Generation To test FADA's applicability beyond text recognition, we apply it to lipogram generation; a constrained generation task where the model must avoid generating a given (forbidden) character (Roush et al., 2022). We prompted Llama-3.1 with 30 diverse topics and applied our proposed guided decoding algorithm (FADA$_{GD}$) by setting the target frequency of the forbidden character to zero (i.e., the model must not generate it). We repeated this process for three different forbidden characters and compared standard beam search with FADA$_{GD}$ decoding, measuring violation rate and perplexity. Results show that FADA effectively enforces character-level constraints during open-ended generation. Full results, topic list, and prompts appear in Appendix E.

Beam size sensitivity study Undertaking a sensitivity study, we investigate the effect of beam size, comparing the performance of the fine-tuned backbone model with and without training-time alignment (i.e., FADA$_{TR}$). We examine beam sizes of 1, 5 and 7. Increasing the beam size in the standard fine-tuned TrOCR model yields only minor gains. In contrast, FADA$_{TR}$ shows stronger improvements, especially at beam size of 1, where it consistently outperforms the baseline even with a beam size of 7. This shows that frequency-aware training boosts transcription quality without relying on large beams, with smaller returns as beam size grows. For more results, see Appendix H.

5 Discussion

Importance of inference-time alignment Table 4 shows that inference-time alignment (FADA$_{GD}$) yields smaller gains compared to training-time alignment (FADA$_{TR}$), which is expected since the latter can update model parameters. However, inference-time alignment offers several practical advantages: it is computationally lightweight, requires no retraining, and is ideal for low-resource scenarios. It also acts as a post-hoc calibration step, often improving results when combined with FADA$_{TR}$ (Table 4), suggesting that frequency-aware scoring can refine outputs even when the underlying model has already been trained to align with frequency statistics. Finally, it supports zero-shot adaptation, as is shown in our Lipogram experiment (§4), where it effectively guides generation away from forbidden characters without any fine-tuning.

Advantages over Temporal/Regional Tagging (TR-TT Baseline) Contextual-conditioning strategies, such as prepending special tokens to indicate conditioning attributes (e.g., language or time) are widely used in recognition tasks (Liu et al., 2020). While effective, these methods control the model in an implicit way that is hard to interpret or adjust. In contrast, by explicitly aligning predicted character distributions with empirical ones, computed from the training data, FADA makes the alignment process interpretable and transparent. Also, it gives control to users over the output, who can modify the target distribution, encouraging (or discouraging) certain characters.

Differences from Domain Adaptation Methods Traditional domain adaptation methods typically assume that a model is trained on a single source-domain (e.g., modern printed text) and then adapted to a different target-domain with limited labeled data, using techniques such as feature alignment or adversarial training. In contrast, FADA is not designed to transfer between domains, but to handle *intra-dataset distributional variation* within a single dataset. Instead of treating each subset (e.g., century, region) as an isolated domain, we train a single model on the entire dataset and guide it to respect the empirical character distribution associated with each training sample. This allows the model to learn general task-relevant representations while adjusting its output to reflect expected patterns of the specific input context without requiring separate adaptation stages or domain boundaries.

Practicality and Feasibility of Domain-specific Character Frequency Priors Both our training-time alignment and inference-time guided decoding methods rely on access to domain-specific intra-dataset character frequency distributions. While such information may not always be available, our focus is on settings where these distributions are relatively stable and can be approximated from external sources. For example, in historical text or regional speech datasets, the set of frequently used characters tends to shift gradually over time or geography, making it feasible to estimate frequency distributions using linguistic studies or domain-specific corpora. We therefore propose FADA as a flexible framework, where alignment can be applied when such priors are available or reasonably inferred, rather than assuming strict access to domain labels at all times.

Applicability to Higher-order Linguistic Patterns Although FADA is formulated as a unigram-level alignment method, it is inherently extensible to higher-order linguistic patterns (e.g. bigrams). One approach involves directly matching predicted n-gram frequency distributions to known domain-specific n-gram priors, encouraging the model to reflect more structured co-occurrence statistics. Alternatively, as we already exploit in our experiments with Transformer-based models, the model may generate subword units, while FADA continues to operate on the decoded character sequence.

Computational Overhead and Reproducibility Training with FADA_{TR} introduces a modest 4–6% increase in training time due to the alignment loss computation, while FADA_{GD} slows decoding by 7–9% due to the added frequency-based scoring. The memory overhead is minimal. Given the accuracy improvements and character-level control, this trade-off remains favorable in most cases. All models were trained with fixed seeds to ensure reproducibility. For CTC-based models, we used the deterministic Baidu CTC implementation. Training/fine-tuning ran for up to 30 epochs with early stopping (patience $= 4$), using a batch size of 4 and a beam size of 5. Details on the λ hyperparameter tuning are included in Appendix G.

6 Conclusion

We present FADA (Frequency-Aware Distribution Alignment), a framework mitigating character relative frequency distribution shifts in text and speech recognition, consisting of two components: (i) a training-time alignment loss minimizing discrepancies between predicted and empirical character distributions; (ii) an inference-time guided decoding frequency-based algorithm that dynamically adjusts predictions. Our experiments on HTR and ASR demonstrate consistent performance improvements across datasets, languages, writing systems, and model architectures. Directions of future work include extending FADA by incorporating bi-grams and tri-grams, enabling alignment at the sequence level rather than for individual characters (unigrams). Also, we plan to explore case-sensitive frequency distributions, and to improve recognition of named entities and specialized terminology. Finally, we aim to develop adaptive weighting strategies for λ, allowing a dynamic adjustment instead of relying on a fixed hyperparameter.

Limitations

While FADA consistently enhances recognition performance across multiple datasets and models, its effectiveness relies on the availability of empirical character frequency distributions. In scenarios where such distributions are not well-defined or highly variable, its impact may be less pronounced. Additionally, the current approach focuses on unigram frequency alignment, leaving room for future exploration of higher-order character dependencies. Although FADA introduces only a modest computational overhead, optimizing its efficiency for real-time applications remains an area of interest. These considerations highlight directions for further refinement rather than fundamental constraints, as FADA remains broadly applicable across different recognition settings.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Anthropic. 2024. Claude 3.5 sonnet.

Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. 2023. NEUROSTRUCTURAL DECODING: Neural text generation with structural constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9496–9510, Toronto, Canada. Association for Computational Linguistics.

Ron Carico, Keaton Kaplan, Kyler Hazelett, Megan Dillon, and Kelly Melvin. 2022. Letter frequency analysis of proprietary prescription drug names in the united states: Minding the zs and qs. *Exploratory Research in Clinical and Social Pharmacy*, 6:100146.

Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Boosting modern and historical handwritten text recognition with deformable convolutions. *International Journal on Document Analysis and Recognition (IJDAR)*, 25:207 – 217.

Anna Chatzipapadopoulou, Ippokratis Pantelidis, Foivos Charalampakos, Marina Samprovalaki, Georgios Moschovis, Panagiotis Kaliosis, Kalliopi V. Dalakleidi, John Pavlopoulos, and Ion Androutsopoulos. 2025. AUEB NLP Group at ImageCLEFmedical Caption 2025. In *CLEF2025 Working Notes*, CEUR Workshop Proceedings, Madrid, Spain. CEUR-WS.org.

Eulrang Cho, Jooyeon Kim, and Hyunwoo J. Kim. 2023. Distribution-aware prompt tuning for vision-language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21947–21956.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille-Levenson, Olivier Brisville-Fertin, Franz Fischer, Michaels Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'Connor, Wouter Haverals, Mike Kestemont, Caroline Vandyck, and Benjamin Kiessling. 2024. CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond. In *2024 International Conference on Document Analysis and Recognition (ICDAR)*, Athens, Greece.

John K. Courtis and Salleh Hassan. 2002. Reading ease of bilingual annual reports. *The Journal of Business Communication (1973)*, 39(4):394–413.

Polina Diurdeva, Elena Mikhailova, and Dmitry Shalymov. 2016. Writer identification based on letter frequency distribution. In *Proceedings of the 19th Conference of Open Innovations Association FRUCT*, FRUCT'19, page 24–30, Helsinki, Uusimaa, FIN. FRUCT Oy.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics.

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gintautas Grigas and Anita Juskeviciene. 2018. Letter frequency analysis of languages using latin alphabet. *International Linguistics Research*, 1:p18.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Wenxin Hou, Jindong Wang, Xu Tan, Tao Qin, and Takahiro Shinozaki. 2021. Cross-domain speech recognition with unsupervised character-level distribution matching. *ArXiv*, abs/2104.07491.

Hang-Rui Hu, Yan Song, Li-Rong Dai, Ian Mcloughlin, and Lin Liu. 2022. Class-aware distribution alignment based unsupervised domain adaptation for speaker verification. In *Interspeech 2022*, pages 3689–3693.

Yijie Hu, Bin Dong, Kaizhu Huang, Lei Ding, Wei Wang, Xiaowei Huang, and Qiu-Feng Wang. 2024. Scene text recognition via dual-path network with shape-driven attention alignment. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(4).

Panagiotis Kaliosis, John Pavlopoulos, Foivos Charalampakos, Georgios Moschovis, and Ion Androutsopoulos. 2024. A data-driven guided decoding mechanism for diagnostic captioning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7450–7466, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Bradley Kjell. 1995. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9:119–124.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI Conference on Artificial Intelligence*.

Yuheng Li, Haotian Liu, Mu Cai, Yijun Li, Eli Shechtman, Zhe Lin, Yong Jae Lee, and Krishna Kumar Singh. 2025. Removing distributional discrepancies in captions improves image-text alignment. In *Computer Vision – ECCV 2024*, pages 405–422, Cham. Springer Nature Switzerland.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Thomas Merriam. 1994. Letter frequency as a discriminator of authors. *Notes and Queries*, 41:467–469.

Meta-AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Marsha Lynn Moreno. 2005. Frequency analysis in light of language innovation: Exploring letter frequencies across time, from the days of old english to the days of now. *Math 187, Spring*.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469.

OpenAI. 2023. Gpt-4 technical report.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2022. Handwritten paleographic Greek text recognition: A century-based approach. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6585–6589, Marseille, France. European Language Resources Association.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Allen Roush, Sanjay Basu, Akshay Moorthy, and Dmitry Dubovoy. 2022. Most language models can be poets too: An AI writing assistant and constrained text generation studio. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 9–15, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Marina Samprovalaki, Anna Chatzipapadopoulou, Georgios Moschovis, Foivos Charalampakos, Panagiotis Kaliosis, John Pavlopoulos, and Ion Androutsopoulos. 2024. AUEB NLP Group at ImageCLEFmedical Caption 2024. In *CLEF2024 Working Notes*, CEUR Workshop Proceedings, Grenoble, France. CEUR-WS.org.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR. In *ICASSP 2023*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2298–2304.

Lazaros Tsochatzidis, Symeon Symeonidis, Alexandros Papazoglou, and Ioannis Pratikakis. 2021. Htr for greek historical handwritten documents. *Journal of Imaging*, 7(12).

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Haoran Wang and Kai Shu. 2025. Make every token count: A systematic survey on decoding methods for foundation models.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Text Style Transfer using Language Models as Discriminators. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik R Narasimhan. 2024. COLLIE: Systematic construction of constrained text generation tasks. In *The Twelfth International Conference on Learning Representations*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A Survey of Controllable Text Generation Using Transformer-Based Pre-Trained Language Models. *Association for Computing Machinery*, 56(3):1–37.

Neil Zhao and Diana Zheng. 2024. An analysis of letter dynamics in the english alphabet. *ArXiv*, abs/2401.15560.

Zhenyu Zhou, Junhui Chen, Namin Wang, Lantian Li, and Dong Wang. 2024. An investigation of distribution alignment in multi-genre speaker recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11596–11600.

A HPGT Dataset - Analysis

This section presents a further analysis on the intra-dataset correlation throughout centuries in the HPGT dataset. In Figure 2, we quantify the correlation trends between character frequency distributions across centuries, computed using Pearson's correlation (r). Adjacent centuries tend to exhibit strong correlations (e.g., 10th–11th: 0.99; 12th–13th: 0.94), whereas distant pairs diverge more substantially (e.g., 10th–16th: 0.91; 13th–15th: 0.89). While correlations above 0.9 generally indicate high similarity, such differences are meaningful in the context of a shared alphabet and may reflect shifts in writing practices, document selection, or biases introduced during data curation. These findings highlight the presence of intra-dataset distribution shifts, whether linguistic or dataset-induced, that can degrade model performance when training ignores subset-specific statistical variation.

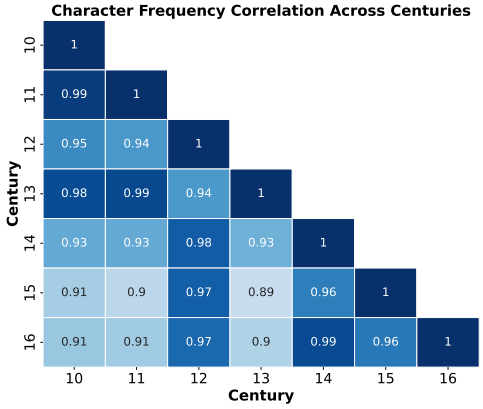

Figure 2: Heatmap of Pearson correlation measuring distributional similarity across centuries; darker shades indicate higher correlation.

B CATMuS Dataset - Analysis

In this section, we examine character frequency distribution shifts in the CATMuS dataset (Clérice et al., 2024), complementing the analysis presented for other datasets. Figure 3 visualizes the relative frequency distributions across four centuries, revealing more subtle intra-dataset variations compared to the rest of the HTR datasets. While some characters exhibit relatively stable frequencies, others show minor fluctuations across different time periods. Figure 4 further illustrates these trends through LOWESS-smoothed frequency distributions, where the black line represents the character distribution for a given century, and the gray lines provide comparisons to the other centuries. The trends indicate that while there are observable shifts, they are not as pronounced as in datasets covering a broader temporal range. Finally, Figure 5 presents a Pearson correlation heatmap, quantifying the similarity between character distributions. The consistently high correlation values (e.g., 15th–16th: 0.99) suggest that distributional shifts are relatively minor, confirming that character frequencies remain largely stable over time in this dataset.

C Synthetic Dataset - Analysis

Similarly, we provide an analysis for our synthetic dataset, which was designed to exhibit controlled character frequency shifts across centuries. Figure 6 presents a heatmap of character usage, reveal-

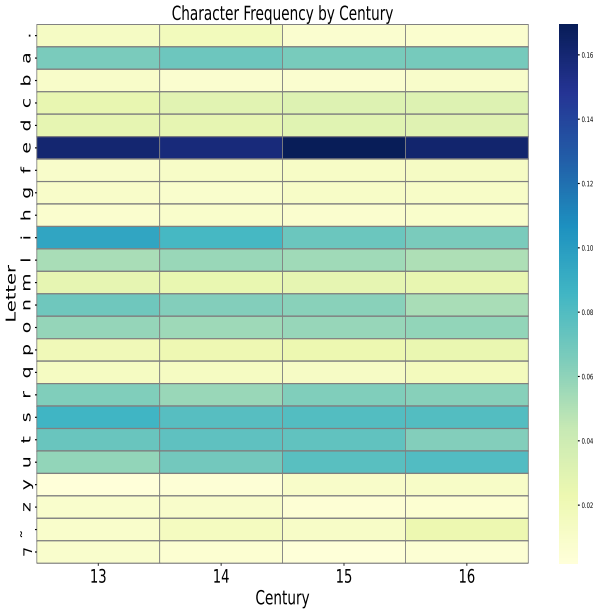

Figure 3: Heatmap illustrating the relative frequency distribution of characters in CATMuS (Clérice et al., 2024).

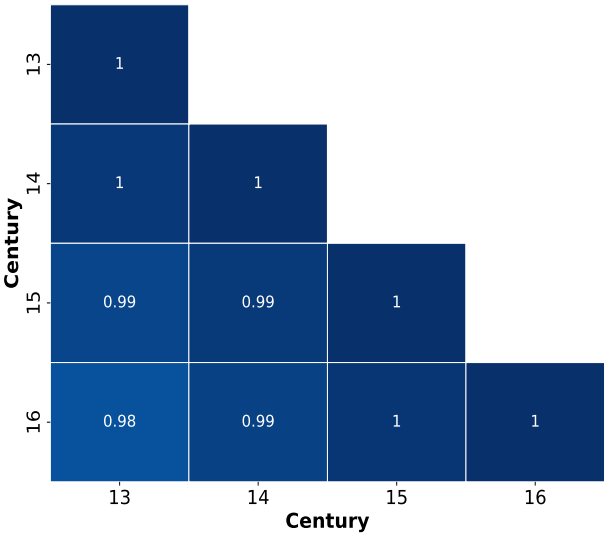

Figure 5: Pearson correlation heatmap of the character frequency distribution across centuries in CATMuS (Clérice et al., 2024). Higher values (darker shades) indicate stronger similarity in character usage, while lower values (lighter shades) suggest greater divergence.

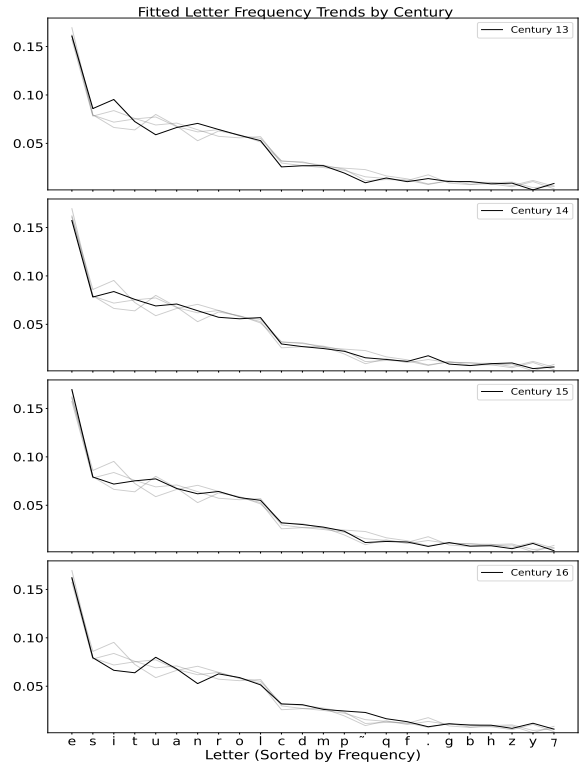

Figure 4: Fitted character frequency trends across different centuries in CATMuS (Clérice et al., 2024). Each subplot represents a specific century, with the black line indicating the LOWESS-smoothed letter frequency distribution for that century. Gray lines represent the character frequency trends of the rest of the centuries for comparison.

ing stable trends for some characters and substantial fluctuations for others.

To further examine these variations, Figure 7 visualizes the fitted frequency distributions per century using LOWESS smoothing. The black line represents the smoothed trend for a given century, while gray lines show distributions from the rest of the centuries for comparison. Lastly, Figure 8 quantifies distributional similarity via a Pearson correlation heatmap. Adjacent centuries show strong correlations (e.g., 11th–12th: 0.97), while more distant ones exhibit lower values (e.g., 10th–12th: 0.89), indicating more pronounced shifts.

D EdAcc Dataset - Analysis

Similarly, we provide an analysis for our ASR dataset. Figure 9 presents a heatmap visualizing the intra-dataset character frequency distributions among the four linguistic backgrounds, exploiting some subtle but noticeable patterns between them. Figure 10 presents the fitted frequency distributions per linguistic background further enhancing the previously stated observations as the black lines deviate compared to the grey lines in each of the four plots.

E Lipogram Generation - Extensive Results

To further assess the impact of FADA_{GD} beyond recognition tasks, we applied it in a controlled text generation setting focused on lipogram con-

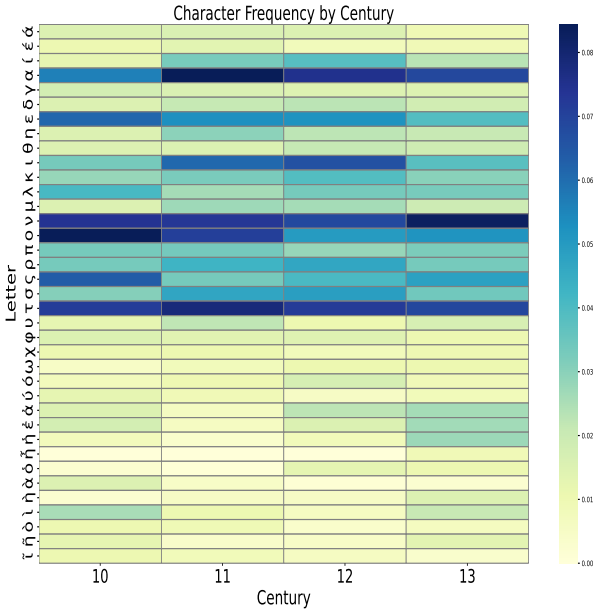

Figure 6: Heatmap illustrating the relative frequency distribution of Greek characters in our synthetic dataset (see §4.1).

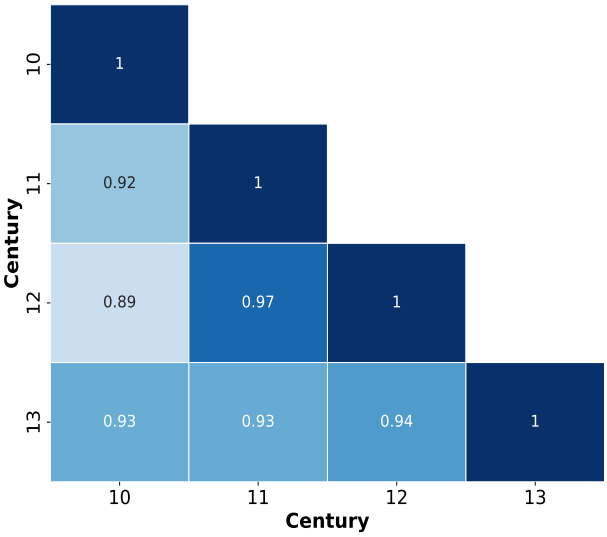

Figure 8: Pearson correlation heatmap of the character frequency distribution across centuries in our synthetic dataset. Higher values (darker shades) indicate stronger similarity in character usage, while lower values (lighter shades) suggest greater divergence.

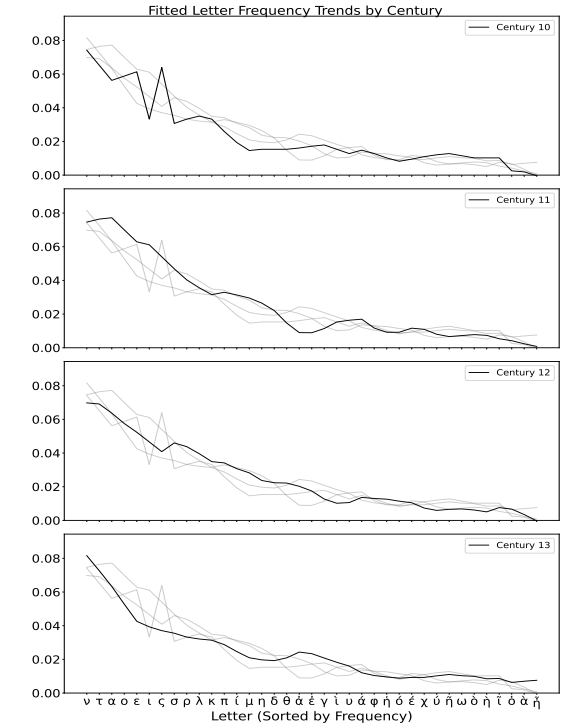

Figure 7: Fitted character frequency trends across different centuries. Each subplot represents a specific century, with the black line indicating the LOWESS-smoothed letter frequency distribution for that century. Gray lines represent the character frequency trends of the rest of the centuries for comparison.

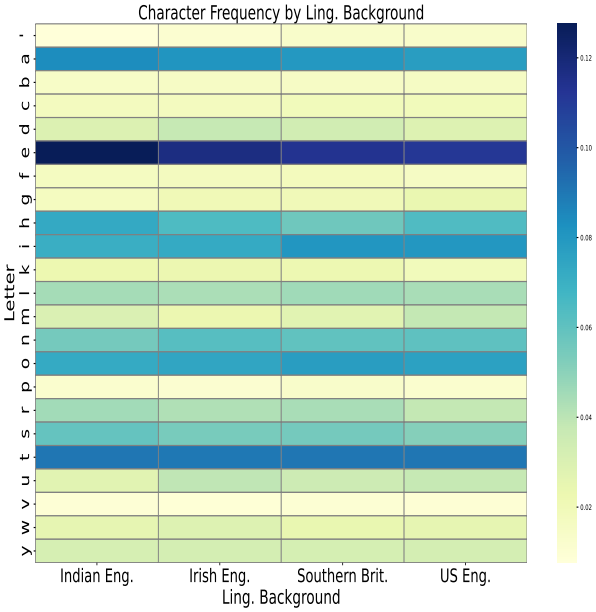

Figure 9: Heatmap illustrating the relative frequency distribution of characters in EdAcc (Sanabria et al., 2023).

straints. In this experiment, we tested whether FADA$_{GD}$ could steer an LLM (`Llama3.1`) to generate coherent passages while avoiding a specific character. The task involved generating short texts (4-6 sentences) on various topics while ensuring that a predetermined character never appeared. This provided a structured way to evaluate how effectivelyFADA$_{GD}$ can influence character distributions in an open-ended autoregressive setting.

We selected three different forbidden characters, namely “A”, “E” and “L”. In each case, we applied FADA$_{GD}$ during decoding by setting the

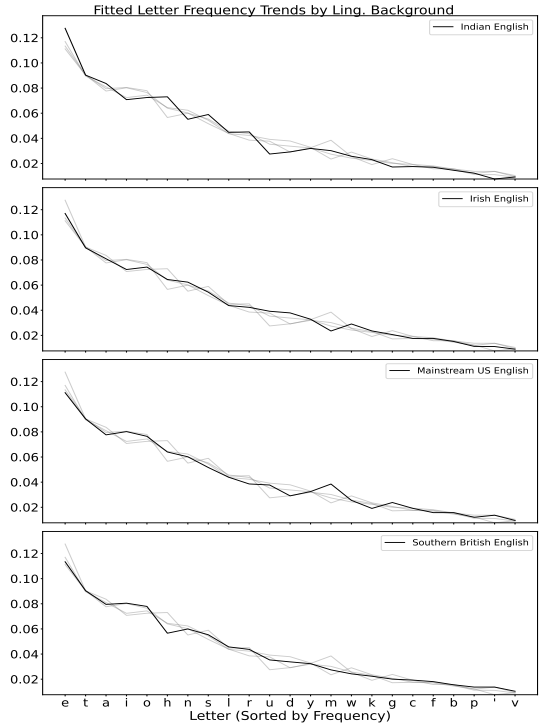

Figure 10: Fitted character frequency trends across different linguistic backgrounds. Each subplot represents a specific background, with the black line indicating the LOWESS-smoothed letter frequency distribution for that background. Gray lines represent the character frequency trends of the rest of the backgrounds for comparison.

forbidden character's target relative frequency to zero.To assess its impact, we compared two decoding configurations: standard beam search, where no constraints were applied, and FADA-GD, which introduced frequency-aware scoring to penalize sequences containing the forbidden character.

For each configuration, the model was prompted with 30 different topics covering diverse subjects. The full list of topics is presented in Table 5, while the prompt is presented in Figure 11. The generated text was evaluated based on the following metrics:

- **Violation Rate**: The percentage of outputs where the forbidden character appeared at least once.
- **Perplexity**: A measure of fluency and coherence in the generated text.
- **Readability**: An assessment of how natural and comprehensible the generated text remains under the applied constraints. We calculated it using the *Flesch Reading Ease Score* (Courtis and Hassan, 2002), which evaluates how easy a text is to read. Higher scores indicate more readable text, with values above 60 considered easy to read, while lower scores (below 30) indicate more complex writing.

We present the numerical results in Table 6. The violation rate succesfully decreases across all three forbidden characters ("A", "E", "L"), with the most notable reduction observed for "E", where the percentage of outputs containing the forbidden character drops from 10.60% to 8.66%, a relative improvement of 18.3%. This confirms that frequency-aware decoding successfully discourages character occurrences without the need for model retraining.

In terms of perplexity, which reflects the fluency and coherence of the generated text, the results show a marginal increase for A and E, while it slightly improves for L. This suggests that enforcing character-level constraints withFADA_{GD} does not significantly degrade text fluency, with differences remaining within acceptable bounds. Readability scores, measured using the Flesch Reading Ease score (Courtis and Hassan, 2002), exhibit improvements for "A" and "L", while remaining nearly unchanged for "E". Overall, these findings highlight FADA_{GD} as an effective approach for guiding language model outputs at the character level while maintaining fluency and readability.
The prompt used in our experiments was:

F Per Century/Region Results

To further assess the impact of FADA across different temporal and linguistic variations, we report per-century and per-region results in Tables 7–10. These results allow us to examine how frequency-aware alignment affects model performance in subsets characterized by distinct character distributions.

For HTR, the results on the HPGTR dataset reveal that FADA provides substantial benefits in centuries where character distributions deviate most from the overall dataset trend. This effect is particularly noticeable in later centuries, where orthographic variations and evolving linguistic conventions introduce additional challenges for recognition models. Similarly, for the CATMuS dataset, where distributional shifts are more subtle, FADA still contributes to performance improvements, particularly in centuries with greater deviations from the dataset-wide character frequencies. The Synthetic dataset further reinforces these observations, as the controlled frequency shifts across different subsets allow us to systematically evaluate the method's effectiveness

ID	Topic
1	A person who went to the post office.
2	A dog at the park.
3	A family that went to the beach.
4	An elephant in the jungle.
5	A person who went to the aquarium.
6	A child who lost a toy at the mall.
7	A bird flying over a quiet village.
8	A person who got stuck in an elevator.
9	A cat exploring a new garden.
10	A teacher who helped a struggling student.
11	A couple hiking in the mountains.
12	A farmer working in the field.
13	A boy who found a hidden cave.
14	A woman who forgot her umbrella on a rainy day.
15	A man fishing by the river.
16	A team that won a soccer match.
17	A child who met their favorite author.
18	A chef preparing a special dish for a celebration.
19	A group of friends camping under the stars.
20	A traveler visiting an ancient temple.
21	A robot learning how to read.
22	A young girl who dreamed of flying.
23	A scientist discovering a new species.
24	A person who got lost in a museum.
25	A dog who saved a child from danger.
26	A boy who built a treehouse with his friends.
27	A person who volunteered at an animal shelter.
28	A musician performing on a busy street.
29	A fisherman who caught a rare fish.
30	A person planting trees in a park.

Table 5: Topics used in the Lipogram Generation experiment.

```
A lipogram is a form of writing where one
or more specific letters are deliberately
avoided throughout the text.
This constraint challenges writers to be
creative and resourceful, often resulting
in unique word choices and phrasing.
Lipograms can be fun and artistic
but also difficult, especially when
the omitted letter is common.
For example, avoiding the letter
'e'—the most frequently used letter
in English—requires significant
effort and skill.
Writers must find alternative ways
to express ideas without breaking
the constraint, which can lead
to inventive language.

TASK:
Generate a short story
(maximum 150 words) about {topic}.
You have to follow a lipogram constraint.
The lipogram constraint is that the letter
{forbidden_character} must not appear
anywhere in the text.
First, go over the task.
Then, explain how you avoided
using the letter {forbidden_character} .

FORMAT:
To ensure clarity and easy readability,
format your output into a JSON.
Use the following format:
{
    "Story": "<Generated_story>"
}

Text:
```

Figure 11: The prompt template used for our lipogram generation experiments

in handling predefined character distribution discrepancies. Across all cases, the most substantial improvements are observed when both training-time and inference-time alignment are applied together, highlighting their complementary nature.

For ASR, the results from the EdAcc dataset demonstrate that FADA adapts well to different linguistic backgrounds, helping the model better handle accent-based variations that lead to systematic shifts in character usage patterns. While training-time alignment consistently enhances performance across most linguistic categories, the impact of inference-time alignment varies. In some cases, guided decoding further refines predictions, while in others, its effect is more limited, suggesting that the degree of benefit depends on the severity of distributional shifts and the alignment between training data and inference-time distributions.

G Tuning λ

To balance transcription accuracy and frequency alignment, we conducted a hyperparameter search over λ values ranging from 0.05 to 0.95 in increments of 0.05. Both training-time (FADA_{TR}) and inference-time (FADA_{GD}) alignment were evaluated on the validation set using CER, WER, and NED as metrics. While no single λ value was optimal across all datasets and models, we observed that values in the 0.3–0.6 range generally led to better overall performance.

Lipogram Generation									
	Violation Rate ↓			Perplexity ↓			Readability ↑		
	A	*E*	*L*	*A*	*E*	*L*	*A*	*E*	*L*
Llama3.1	6.48	10.60	3.02	**3.64**	**1.93**	2.24	72.40	**78.97**	79.20
Llama3.1 + FADA$_{GD}$	**5.73**	**8.66**	**2.50**	3.81	1.98	**2.03**	**75.61**	78.11	**79.37**

Table 6: The performance of FADA$_{GD}$ framework in the lipogram generation task.

HPGTR Dataset - Per Century Results (CER)							
	TrOCR-base (Li et al., 2021)						
	10	*11*	*12*	*13*	*14*	*15*	*16*
Standard FT	8.99	15.31	27.18	32.47	50.06	32.37	43.78
FADA$_{GD}$	8.44	14.74	27.27	30.76	49.34	31.61	46.15
FADA$_{TR}$	8.07	12.64	25.48	**27.98**	46.57	27.32	42.09
FADA$_{TR\text{-}GD}$	**8.04**	**12.07**	**25.10**	28.11	**45.82**	**27.15**	**41.36**
	C-RNN (Shi et al., 2015)						
	10	*11*	*12*	*13*	*14*	*15*	*16*
Standard FT	10.26	16.11	27.19	38.78	37.78	31.41	38.54
FADA$_{GD}$	10.20	16.07	27.08	38.14	37.11	30.55	38.51
FADA$_{TR}$	10.13	15.88	26.31	37.84	**35.75**	**29.54**	38.76
FADA$_{TR\text{-}GD}$	**10.05**	**15.77**	**26.27**	**37.79**	35.77	29.59	**38.47**

Table 7: The per-century performance (in terms of CER) of TrOCR-base (Li et al., 2021) and C-RNN (Shi et al., 2015) on the HPGTR dataset (Platanou et al., 2022). Baseline denotes the vanilla model, FADA$_{GD}$ and FADA$_{TR}$ denote the proposed inference-time and training-time alignment methods respectively, while FADA$_{TR\text{-}GD}$ indicates the combination of them.

CATMuS Dataset - Per Century Results (CER)				
	TrOCR-base (Li et al., 2021)			
	13	*14*	*15*	*16*
Standard FT	12.28	**12.07**	9.47	5.76
FADA$_{GD}$	12.11	12.16	9.46	5.79
FADA$_{TR}$	**11.95**	12.16	9.42	4.94
FADA$_{TR\text{-}GD}$	12.10	12.17	**9.18**	**4.90**
	C-RNN (Shi et al., 2015)			
	13	*14*	*15*	*16*
Standard FT	20.34	23.01	15.79	8.92
FADA$_{GD}$	20.11	22.89	15.77	8.73
FADA$_{TR}$	19.87	21.36	15.89	8.67
FADA$_{TR\text{-}GD}$	**19.78**	**21.20**	**15.64**	**8.65**

Table 8: The per-century performance (in terms of CER) of TrOCR-base (Li et al., 2021) and C-RNN (Shi et al., 2015) on the CATMuS dataset (Clérice et al., 2024). Standard FT denotes the fine-tuned backbone model, FADA$_{GD}$ and FADA$_{TR}$ denote the proposed inference-time and training-time alignment methods respectively, while FADA$_{TR\text{-}GD}$ indicates the combination of them.

Synthetic Dataset - Per Category Results (CER)				
	TrOCR-base (Li et al., 2021)			
	GPT-4	*Llama*	*Gemini*	*Claude*
Standard FT	3.96	9.83	5.38	4.43
FADA$_{GD}$	20.21	22.78	5.23	4.39
FADA$_{TR}$	**2.96**	8.87	5.05	4.27
FADA$_{TR\text{-}GD}$	3.23	**8.19**	**4.65**	**3.81**
	C-RNN (Shi et al., 2015)			
	GPT-4	*Llama*	*Gemini*	*Claude*
Standard FT	9.29	14.77	11.32	8.97
FADA$_{GD}$	9.17	13.29	10.88	8.45
FADA$_{TR}$	8.07	11.67	9.16	5.41
FADA$_{TR\text{-}GD}$	**8.03**	**11.48**	**9.12**	**5.21**

Table 9: The per-century performance (in terms of CER) of TrOCR-base (Li et al., 2021) and C-RNN (Shi et al., 2015) on our synthetic dataset. Standard FT denotes the fine-tuned backbone model, FADA$_{GD}$ and FADA$_{TR}$ denote the proposed inference-time and training-time alignment methods respectively, while FADA$_{TR\text{-}GD}$ indicates the combination of them.

Synthetic Dataset - Per Category Results (CER)				
	Whisper-small (Radford et al., 2023)			
	Southern British English	*Indian English*	*Irish English*	*US English*
Standard FT	37.23	25.85	28.37	28.70
FADA$_{GD}$	36.26	25.76	29.51	28.38
FADA$_{TR}$	35.71	25.93	27.37	27.62
FADA$_{TR\text{-}GD}$	37.05	26.17	27.92	27.74

Table 10: The per linguistic background performance (in terms of CER) of Whisper-small on the EdAcc dataset (Sanabria et al., 2023). Standard FT denotes the vanilla model, FADA$_{GD}$ and FADA$_{TR}$ denote the proposed inference-time and training-time alignment methods respectively, while FADA$_{TR\text{-}GD}$ indicates the combination of them.

H Ablation on beam size

As an ablation study, we investigate the effect of the beam size in FADA. To this end, we compare the performance of the standard fine-tuned model with standard beam search against FADA$_{TR}$. We explore beam sizes of 1 (greedy decoding), 5 and 7. For both datasets (HPGTR and Synthetic), increasing the beam size in the fine-tuned TrOCR model yields only marginal improvements, particularly in CER and WER. We present the results in Table 11. While a larger beam (e.g., 5 or 7) generally reduces errors slightly compared to greedy decoding (beam size = 1), the gains are relatively small. This suggests that the standard beam search does not significantly improve recognition performance beyond a certain threshold.

Beam Size Ablation Results						
Model	**CER ↓**			**WER ↓**		
	1	*5*	*7*	*1*	*5*	*7*
HPGTR						
Standard FT	29.01	26.92	26.87	76.25	73.82	73.31
FADA$_{TR}$	**25.33**	**25.16**	**24.73**	**70.41**	**70.47**	**69.98**
Synthetic Dataset						
Standard FT	6.44	5.92	5.91	18.98	17.89	17.86
FADA$_{TR}$	**5.46**	**5.30**	**5.28**	**17.10**	**16.86**	**16.85**

Table 11: **Beam size ablation on HPGTR and Synthetic datasets.** We report CER and WER for beam sizes 1, 5, and 7.

When applying FADA$_{TR}$, the improvements are more substantial, particularly for beam size = 1, where it consistently outperforms the fine-tuned model even at beam size = 7. This demonstrates that the proposed training-time alignment enhances the model's inherent ability to generate more accurate transcriptions without relying heavily on an extensive search space. Notably, even with beam size = 1, FADA$_{TR}$ achieves lower CER and WER than the fine-tuned model with beam size = 7, highlighting the effectiveness of frequency-aware training. Moreover, FADATR also benefits from a moderate increase in beam size, but the relative improvements diminish as the beam size grows.