Zero-Shot Generative Segmentation of Historical Maps via Natural Language-Driven Vision-Language Models

Jörg Baumann[1][0009−0003−9649−0744], David Fleischhacker[2][0009−0003−4787−9091], Wolfgang Göderle[2][0000−0002−9417−5316], and Roman Kern[1][0000−0003−0202−6100]

[1] Institute of Machine Learning and Neural Computation, Graz University of Technology, Austria {jbaumann, rkern}@tugraz.at
[2] Department of Digital Humanities, University of Graz, Austria {david.fleischhacker, wolfgang.goederle}@uni-graz.at

Abstract. Historical cartographic archives represent a critical yet difficult-to-access resource for spatial analysis, and they continue to pose challenges for systematic machine-readable exploitation. This paper evaluates the potential of novel Vision-Language Models (VLM) as a foundation for the generative segmentation of historical map linework. Following a multi-tiered experimental framework, we compare zero-shot, ensemble-based, and fine-tuned configurations across three diverse historical corpora: the Ferraris Map, Napoleonic Land Registry, and Franciscean Cadastre. Our analysis finds that VLM's ability in generative synthesis as a means of pre-processing effectively lowers the algorithmic barrier for line segmentation, allowing simple intensity thresholding to outperform sophisticated deep-learning detectors. This is substantiated by F_1 scores of 42.47, 66.47, and 61.37 across the Ferraris, Napoleonic, and Franciscean corpora respectively. The best performance is consistently achieved by an end-to-end ensemble approach, with architectures such as Qwen 2511 and Flux.2 establishing F_1 scores of 55.24, 73.66, and 75.45. While these generative models have been found to lack the local pixel-perfect precision of discriminative baselines, our findings suggest superior topological awareness and structural continuity, particularly on highly complex archival map data. Furthermore, parameter-efficient fine-tuning via LoRA increases generative stability and reduces variance, but it remains prone to systematic biases not found in the zero-shot ensemble approach. Our findings validate VLM-driven synthesis as a natural language-guided zero-shot paradigm for cartographic analysis. By significantly reducing the requirement for extensive manually annotated training data, this framework offers a highly accessible and generalizable solution that facilitates the transformation of diverse archival records into structured geospatial data.

Keywords: Zero-Shot · Low-Rank Adaptation · Vision-Language Models · Line Segmentation · Historical Cartography · Historical Documents

1 Introduction

Historical maps constitute a rich source for historical and archaeological research. Unlike textual records, they encode spatial information systematically and at scale, capturing the configuration of settlements, land use, infrastructure, and natural features across large geographic extents and over extended time periods. However, while textual archives are now routinely transcribed and semantically indexed at scale, cartographic records continue to impede systematic transformation into structured, machine-readable data. Computer vision approaches often struggle with the stylistic heterogeneity of these archives, where models trained on one series, such as polychromatic cadastral surveys, generalise poorly to monochromatic military topographies. Furthermore, physical degradation and the high cost of manual annotation create a significant barrier for interdisciplinary research.
Recent advancements have mitigated this bottleneck through few-shot learning and parameter-efficient fine-tuning of large foundation models, substantially reducing (though not eliminating) the requirement for labelled training data. In this paper, we propose a novel training-free alternative, investigating the capacity of contemporary Vision Language Models (VLMs) to perform line segmentation in a fully zero-shot setting, guided by natural language instructions alone.

We evaluate the Flux and Qwen model families as both a pre-processing layer for established pipelines and as end-to-end generative segmentation engines. Our study spans three stylistically and structurally distinct map corpora: The Austrian Franciscean Cadastre, the Napoleonic Land Registry, and the Ferraris Map. Beyond zero-shot inference, we assess performance gains achievable through Low-Rank Adaptation (LoRA) [10] on a single annotated sheet, investigating how minimal supervision interacts with models' pre-existing semantic capabilities.

2 Related Work

The extraction of information from historical digitized maps has been an active field of research for several decades [7]. Early segmentation approaches focused primarily on techniques such as histogram thresholding [9], grouping pixels by colour similarity [5] or pattern recognition utilizing pre-defined geometrical attributes [16]. More recently, Convolutional Neural Network (CNN) based approaches have displayed strong segmentation capabilities on historical map data [4]. Heitzler et al. [8] successfully performed the segmentation of buildings from the Topographic Atlas of Switzerland following a U-Net [20] based approach. A prevalent challenge in the development of large-scale machine learning models is the acute scarcity of annotated training data for historical map corpora. To address this bottleneck, transfer learning can be applied: models are pre-trained on large, generic datasets and progressively fine-tuned to the target domain using a relatively small set of labelled data. Despite the potentially significant

differences in source and target domain, this approach can lead to improved performance [30]. Oliveira et al. [1] leveraged a pre-trained ResNet encoder to segment Napoleonic cadastral maps of Venice while Petitpierre et al. [17] followed a similar approach in their segmentation of historical maps of Paris.

Xia et al. [28] utilize the strong zero-shot performance of SAM [11] to develop an efficient fine-tuning strategy for historical map segmentation called MapSAM. It leverages weight-decomposed low-rank adaptation (DoRA) [14] to bridge the domain gap between natural imagery and historical maps. Automated prompt generation further decreases the need for extensive manual labelling, narrowing the performance gap relative to a fully supervised U-Net baseline. Sterzinger et al. [22] build on this approach employing linear probing of frozen vision foundation models, such as DINOv2 and RADIO. By exploiting the rich semantic embeddings already present in these models, they demonstrate that segmentation of complex historical map features can be achieved even with very little annotated data. While these models excel at discrete feature extraction, maintaining global coherence across large-scale documents remains a challenge. This was addressed by MapSAM2 [27], which exploits the temporal memory-attention mechanisms of SAM 2 [19]. It treats adjacent map tiles as pseudo video sequence in order to improve contextual consistency across large map sheets.

3 Dataset

To evaluate the robustness and generalization capabilities of our approach, we record performance across three distinct map corpora.

(a) Ferraris (b) Napoleonic (c) Franciscean

Fig. 1. Example images for each of our selected map corpora: The Ferraris Map (a), the Napoleonic Land Registry (b) and the Franciscean Cadastre (c).

3.1 Selection Rationale

The selection of these three corpora was driven by several strategic considerations designed to evaluate the robustness and generalization of our VLM based approach. All selected maps are publicly accessible in high-resolution enabling reproducibility and future benchmarking. Beyond accessibility, these corpora cover a diverse spectrum of European cartographic traditions, each posing individual technical obstacles to automated segmentation. The Ferraris map establishes a baseline for semantic reasoning, where the model must identify rows of vegetation functioning as parcel boundaries, a task beyond simple edge detection logic. In contrast, the Napoleonic registry shifts the challenge towards geometric interference. Here the model must distinguish structural boundaries from a dense, intersecting network of polychromatic survey lines and markings that occlude the target features. Finally, the Franciscean Cadastre, characterized by dense calligraphic and symbolic annotations, presents a unique segmentation challenge requiring the model to leverage high-level semantic disambiguation to restore and maintain boundary continuity. Collectively, this multi-corpus approach ensures that the evaluation is not limited to a single stylistic domain, but instead tests the capacity of VLM models to perform instruction based reconstruction under a wide range of historical and technical conditions.

3.2 Map Corpora

Ferraris Map Featuring a total of 275 sheets of a scale of 1:11,520, the Ferraris map covers the entirety of modern day Belgium and Luxembourg, the southern Netherlands as well as parts of Germany. Driven by the request for a military survey, it was created between 1770 and 1778 and represents the first large scale mapping of Belgium and parts of Western Europe [24]. Capturing the state of land usage shortly before the beginning of the Industrial Revolution, it represents an important source of information for a variety of interdisciplinary fields. It was drawn in full colour, employing symbols and text to provide a detailed depiction of features such as land cover, roads, waterways and buildings. Notably, rows of vegetation symbols such as trees or shrubbery frequently serve in place of solid lines to delineated parcel boundaries. Structural outlines are often characterized by discontinuities, while navigational guidance lines appear as low contrast black markings. Furthermore, background colours generally appear washed displaying a large range of variance.

Napoleonic Land Registry Initiated in 1807, the French parcel cadastre was designed to systematically document land ownership for optimized fiscal administration. Approximately 40,000 communes were surveyed at scales ranging from 1:1,250 to 1:2,500 with the project reaching completion in mainland France by 1850. The maps depict individual plots identified by unique numerical labels and classified by land use such as meadows, vineyards, fields and similar. While built structures are typically depicted as red or yellow geometric shapes and waterways utilize blue fillings, the majority of the sheets are monochromatic.

However, the structural boundaries are frequently occluded by red, yellow, or faint black survey guidance lines and large angled cross symbols marking triangulation points. Ink stains, paper deterioration and calligraphic annotations further obscure boundary lines thereby increasing the segmentation challenge.

Franciscean Cadastre Commissioned between 1817 and 1861, the Franciscean Cadastre constitutes the most comprehensive documentation of land ownership within the Habsburg monarchy, encompassing vast parts of Central Europe [18]. The corpus is renowned for its rigorous systematic execution, achieving a scale of up to 1:720 in urban centres while transitioning to 1:5,760 in remote mountainous regions. Typically for cadastral maps, the topography is partitioned into individual parcels, each identified by a unique numerical designation and clearly delineated by distinct boundaries. The maps are polychromatic, utilizing area fills to denote parcel classification and usage. These fills frequently spill across boundaries into adjacent parcels, effectively obscuring smaller structures and complicating accurate segmentation. Furthermore, the sheets are densely populated with symbolic markers to enrich the representation of land features, such as specific kinds of trees in forested zones or vine symbols for viticulture. These often overlap with the structural boundaries of roads, waterways, and buildings. This high symbolic density necessitates a model capable of distinguishing between structural geometries and superimposed cartographic icons.

4 Methodology

We employ a multi-tiered framework designed to evaluate the utility of VLM integrated generative models across varying degrees of task specialisation. The experimental setup is structured into three distinct phases: first, we assess the capabilities of these architectures as a zero-shot pre-processing layer to enhance existing segmentation pipelines. Second, we evaluate their capacity to function as end-to-end generative segmentation engines. And third, we analyse the performance gains achieved through task-specific fine-tuning on a single historical map corpus.

4.1 Model Selection and Architectural Diversity

To prioritize transparency and reproducibility, our experimental setup exclusively utilizes open-weight models readily available for local deployment. We specifically examine the Flux and Qwen model families, which represent two dominant architectural paradigms in current visual synthesis: Rectified Flow Matching [13] [15] and Multimodal Diffusion Transformers [6]. Both families are fundamentally VLM-driven, utilizing high capacity vision language encoders to map semantic instructions to precise spatial reconstructions. Our selection encompasses a diverse range of architectural scales and precision levels to assess the correlation between model capacity and cartographic segmentation accuracy. Within the Flux ecosystem, we evaluate the 12B Flux.1 Kontext [dev] [12], the

32B Flux.2 [dev] [2] utilzing FP8 mixed precision, and the compact 9B Flux.2 [klein] Base. Within the Qwen family [26], we compare the 20B Qwen Image Edit (FP8) and its successor, 20B Qwen Image Edit 2511 (Q8).

4.2 Prompt Engineering

To ensure a fair comparison across the diverse models in our study, we adopted a standardized, time-constrained approach to prompt development. Each model was allotted a maximum of one hour for iterative refinement to simulate real-world deployment constraints and prevent disproportionate optimization. As observed in contemporary reasoning-heavy Large Language Models (LLMs), the framing of an instruction is a primary determinant of zero-shot performance [25]. Our experiments suggest that VLM-driven synthesis models exhibit a similar sensitivity, where the model's capacity for spatial reconstruction is dependent on the semantic clarity of the prompt. We identified two primary archetypes for successful task guidance in the context of historical cartography. The first paradigm involves a structured instructional specification. It is characterised by a detailed description of the task objectives, a sequential outline of the operations required for execution, and a precise definition of the target semantic output. Conversely, the second approach utilizes an imperative structure kept short and to the point, consisting of concise, often numbered instructions that rely on direct and unambiguous language to guide the model. Examples of both paradigms are provided in Listings C.1 and C.2. While the more comprehensive narrative prompts achieved superior performance with larger, current generation models such as Flux.2 [dev] and Qwen Image Edit 2511, the imperative, step oriented style proved more effective for older or smaller architectures like Flux.1 Kontext [dev].

4.3 Generative Pre-Processing Layer

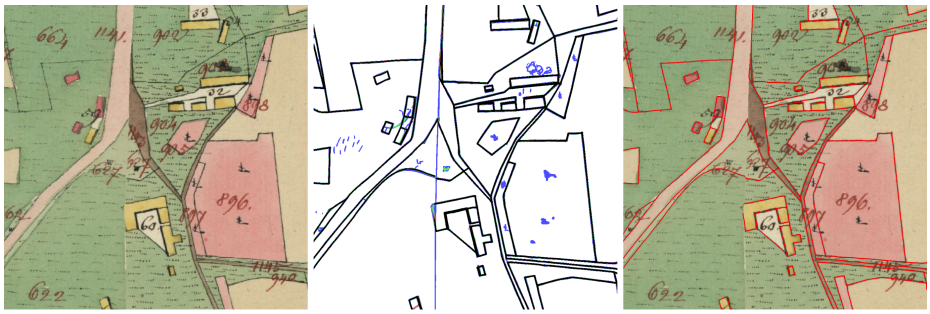
To evaluate VLM-driven synthesis capabilities employed as a pre-processing component, we task each model with generating a canonical reconstruction of the source map tiles (see Figure 2). This canonical version is defined as a semantically filtered image where all non-structural elements such as written annotations, cartographic symbols, and background textures are removed alongside any archival degradations such as ink bleeds. Boundary lines previously occluded by such elements are fully restored. The resulting canonical output is subsequently processed through a standardized intensity-based binarization pipeline. This involves converting the model generated output to greyscale and applying a thresholding operation to isolate the primary structural boundaries. To mitigate the impact of generative stochasticity and potential hallucinations inherent in these architectures, we implement a multiple inference sampling strategy, executing each prompt ten times per dataset entry. These ten iterations are then individually processed through the binarization pipeline and evaluated against the manually annotated ground truth.

Fig. 2. Flux.2 generative outputs across three map corpora. (a) Source image, (b) ground truth, (c) canonical synthesis, (d) thresholded mask, and (e) end-to-end ensemble. Rows (top-to-bottom): Ferraris, Napoleonic, Franciscean datasets.

4.4 End-to-end Synthesis for Generative Line Segmentation

In the second phase of our experiment, we investigate the capacity of VLM-driven synthesis models to function as standalone segmentation engines, tasked with the direct generation of binary boundary masks (see Figure 2). Unlike the multi-stage pre-processing workflow, this approach circumvents traditional intensity-based thresholding. It necessitates that the VLM-driven models execute both semantic disambiguation and pixel-level classification within a single generative pass. The synthesized output is binarized using a fixed threshold of 127 to eliminate low-intensity artifacts. To mitigate the impact of localized hallucinations and maintain structural integrity, we implement a pixel-wise ensemble strategy (see Figure 3). Each prompt is executed ten times per map tile. Preliminary empirical analysis indicated that a ten-pass iteration provides a stable ensemble, beyond which marginal performance gains were offset by diminishing returns in computational efficiency. The results are then aggregated into a unified output via majority voting. A pixel is classified as structural boundary in the final ensemble mask only if it is predicted as such by at least 60% (i.e., ≥ 6 out of 10) of the model's outputs. While the pre-processing evaluation analyses the individual variance of the VLM as a component, the end-to-end evaluation employs a majority-vote to achieve ensemble driven stability (see Table 5). This ensures that the final binary output reflects the model's consistent semantic in-

terpretations while filtering out transient generative noise that may occur in isolated inference passes.



(a) Input (b) Ensemble Votes (c) Segmentation

Fig. 3. Example of the end-to-end generative ensemble pipeline. (a) Input tile, (b) segmentation agreement over 10 iterations, colour coded by vote frequency: 0–2 votes (blue), 3–4 (green), 5 (red), and ≥ 6 (black). (c) The finalized ensemble segmentation overlaid on the original source image.

4.5 Fine-tuning and Domain Adaption

To surpass the constraints of zero-shot inference, we finally investigate the performance gains achievable through supervised fine-tuning for end-to-end line segmentation. For this phase, Qwen 2511 Image Edit and Flux.2 [dev] were selected due to their strong consistent performance across all three datasets (see Table 1). We employed Low Rank Adaption (LoRA) [10] to perform parameter-efficient fine-tuning, allowing the models to internalize specific cartographic features without compromising its foundational vision-language capabilities. The Franciscean Cadastre was chosen as the primary adaption corpus due to its high-resolution availability, its high annotation density, and its diverse, polychromatic style, which represent a significant segmentation challenge. The training set was derived from a single, hand annotated map sheet, partitioned into 117 tiles of 1024×1024 pixels with a 256 pixel overlap along both axes to ensure spatial continuity. Training was conducted for 5000 steps on an NVIDIA H200 GPU. To prevent any data leakage, we strictly isolated the training sheet from the sheets used for evaluation.

5 Evaluation

To quantify the performance of our VLM-driven generative approaches, we assembled an evaluation set for each of the three map corpora described in Section

3. Each evaluation set consists of 20 representative map tiles of 1024×1024 pixels, paired with manually annotated binary ground truth masks. These masks contain all structural boundaries, encompassing both the explicit geometric lines found in the Napoleonic and Franciscean datasets as well as the semantic vegetation based boundaries prevalent in the Ferraris corpus (see Figure 1). Performance is measured using two primary metrics: the F_1 score, to evaluate the harmonic mean of precision and recall, and the mean Intersection over Union (IoU), to asses the spatial overlap accuracy between ground truth and prediction.

5.1 Threshold Optimization: ODS

Our baseline models generate greyscale probability maps with pixel intensities ranging from 0 to 255. To derive a binary segmentation from these probabilistic outputs, as well as for the thresholding step of the generated canonical map image, an optimal threshold must be determined. We perform an exhaustive search across 99 evenly spaced thresholds within the range $[1, 254]$, ranking them based on their achieved F_1 scores. From this procedure, we determine the Optimal Dataset Scale (ODS), representing the single threshold that maximizes performance across the entire corpus. This value is then used for all further evaluations and ensures a fair comparison between the predictions of our end-to-end VLM ensemble and the probabilistic outputs of our baseline edge detection frameworks.

5.2 Spatial Tolerance and Ground Truth Dilation

A significant challenge in the quantitative evaluation of historical map segmentation is the potential discrepancy between generative model outputs and manual annotations. The original cartographic linework exhibits variable stroke widths of fluctuating intensities, frequently made worse by archival degradation or overpainting during the map's creation. To account for this variance and the inherent subjectivity of the manual labelling process, we incorporate a measure of spatial tolerance through morphological dilation. We apply a disk-shaped structuring element to the binary ground truth masks across a range of radii $r \in 0, 1, 2, 3$ pixels. The $r = 0$ case represents a strict pixel perfect match while increasing radii provide a larger measure of tolerance accounting for minor spatial misalignments. This multi-scale approach ensures that the evaluation captures the structural fidelity of the predicted boundaries without penalizing the model for uncertainties inherent in the source material or variances introduced during the manual annotation process. Thus, by accounting for both the physical degradation of the cartographic ink and the inevitable pixel-level subjectivity of human labelling, we provide a more robust assessment of the model's architectural performance.

5.3 Baselines

To establish a thorough performance benchmark, we compare our VLM-driven approach against four representative edge detection methods that span the

methodological spectrum from classical signal processing to modern generative modelling (see Figure 4). We utilize the classical Canny operator [3] to establish a fundamental performance floor for the segmentation task. To evaluate the effectiveness of specialized discriminative architectures, we include PiDiNet [23] and TEED [21]. PiDiNet uses Pixel Difference Convolutions to efficiently capture fine-grained local structures with minimal parameter overhead, while TEED represents a recent advancement in parameter-efficient CNNs specifically optimized for thin edge detection without the need for extensive pre-training. Furthermore, we benchmark our results against Diffusion Edge [29], a state-of-the-art generative framework. Since it also employs a stochastic reconstruction process, it serves as an important point of comparison to analyse the performance impact provided by semantic and linguistic reasoning capabilities over purely vision based generative models. Finally, to establish a task specific benchmark for the end-to-end generative segmentation, we implement the few-shot linear-probing approach by Sterzinger et al. [22]. This method was selected due to its demonstrated strong performance in the original evaluations, where it was established as a robust approach for cartographic feature extraction. Following the protocol established in the Siegfried benchmark dataset [28], we utilize 10 training and 8 validation tiles of 224×224 pixels for this few-shot baseline.

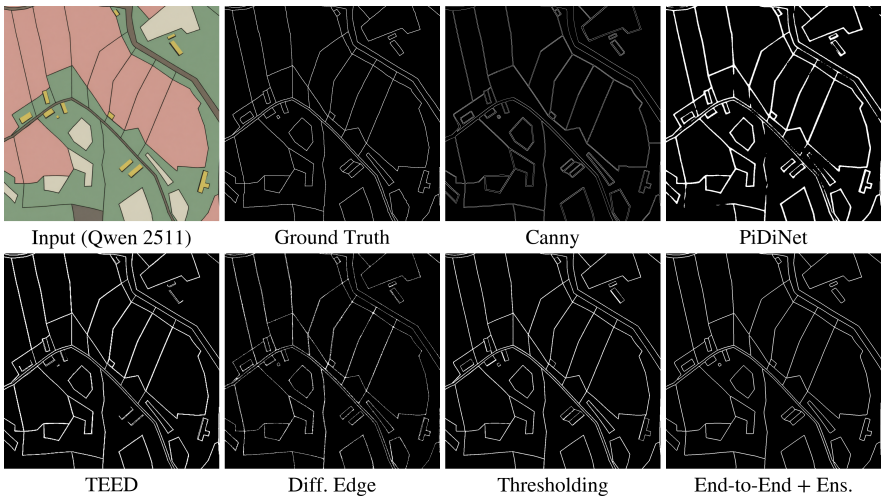


Fig. 4. Visualisation of baseline and thresholding approaches applied to a Qwen 2511 generated canonical image, and the corresponding VLM end-to-end output.

6 Results

Performance metrics for the three map corpora are summarized below, comparing zero-shot VLM configurations, canonical synthesis and end-to-end

segmentation, against five baselines. Figures 2 and 3 illustrate the visual outputs and ensemble behaviour underpinning this evaluation. Figure 4 contrasts our end-to-end pipeline against the baselines on the Franciscean corpus. Strict spatial correspondence (r = 0) is documented in Table 1, while generative stability and stochastic characteristics are assessed in Figure 5 and Table 3 (for exhaustive data, see Tables A.6 and A.7). VLM inference times are listed in Table 2. Table 4 evaluates our VLM-driven architectures against the few-shot baseline across the full range of spatial tolerance radii. Finally, the performance trajectory across progressively increasing spatial tolerance is illustrated through Figure 6.

Table 1. Performance metrics (r=0) across all evaluation corpora. Comparison of canonical VLM generation and end-to-end ensemble. The first row per dataset denotes baseline performance on raw input tiles.

	Canny		**PiDiNet**		**TEED**		**Diff. Edge**		**Thresh.**		**End-to-end**	
Model	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU
Ferraris Map												
Original	13.28	7.19	24.81	14.27	22.05	12.71	20.85	12.05	34.67	21.73	–	–
Flux.1	22.03	12.65	26.77	15.89	29.56	18.00	26.80	16.19	30.24	18.83	50.65	34.67
Flux.2	24.09	13.87	30.98	18.52	34.46	21.17	32.32	19.60	42.47	27.66	54.23	38.10
Flux.2 klein	21.17	11.97	27.12	15.82	29.13	17.31	28.43	16.90	38.78	24.56	50.01	34.04
Qwen	18.87	10.56	26.41	15.42	27.08	16.50	23.83	14.23	37.65	24.44	46.34	30.97
Qwen 2511	20.75	11.71	28.72	16.94	31.02	18.91	24.92	14.68	37.16	23.69	**55.24**	**38.84**
Napoleonic Land Registry												
Original	23.01	13.14	37.15	23.07	58.99	42.28	44.63	28.87	51.13	35.06	–	–
Flux.1	27.00	15.90	45.33	29.89	63.08	47.89	54.67	38.72	66.03	50.86	71.64	56.55
Flux.2	22.28	12.80	44.85	29.24	53.38	37.58	45.06	30.01	66.47	50.75	72.01	57.01
Flux.2 klein	14.80	8.27	34.07	20.94	37.00	23.46	28.06	17.65	59.47	43.16	69.28	53.65
Qwen	24.06	14.18	36.85	23.90	47.77	35.06	39.44	26.74	45.99	33.68	6.14	3.37
Qwen 2511	31.86	19.37	49.40	33.79	61.53	46.84	52.28	36.90	60.84	45.45	**73.66**	**59.09**
Franciscean Cadastre												
Original	16.54	9.14	35.45	21.89	41.78	27.02	42.21	26.98	33.70	21.25	–	–
Flux.1	23.73	13.74	43.41	28.35	54.20	38.59	53.83	38.05	56.33	41.30	70.31	55.51
Flux.2	5.09	2.66	27.54	16.22	27.34	16.10	12.46	6.92	56.16	40.13	75.03	60.23
Flux.2 klein	12.57	6.84	35.62	21.92	37.77	23.68	26.52	15.99	61.37	44.90	70.86	55.09
Qwen	25.14	14.66	43.90	28.91	56.26	41.20	52.24	36.81	53.68	39.44	55.84	41.70
Qwen 2511	16.97	9.57	34.25	21.44	37.41	24.38	33.26	21.27	49.89	35.08	**75.45**	**60.93**

Table 2. Mean inference times for canonical synthesis and end-to-end VLM tasks for a single pass of an image.

Model	Time / Image (s)
Flux.1	17.89
Flux.2	49.93
Flux.2 klein	34.12
Qwen	45.54
Qwen 2511	58.14

Table 3. End-to-end segmentation performance for r=0 without ensemble showing the best and worst results across the Franciscean corpus.

model	F_1	Std	Best	Worst	IoU	Std	Best	Worst
Flux.1	60.10	16.66	76.48	0.62	44.62	14.18	61.92	0.31
Flux.2	**68.89**	**7.12**	81.28	35.82	**52.96**	**7.76**	68.47	21.82
Flux.2 klein	60.15	8.46	75.41	**37.93**	43.52	8.58	60.53	**23.40**
Qwen 2511	59.93	16.70	**81.39**	0.00	44.54	14.95	**68.62**	0.00
Qwen	38.98	25.16	78.52	2.31	27.30	19.98	64.63	1.17

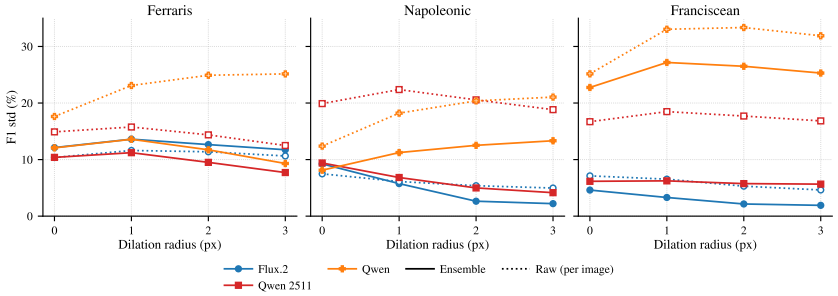


Fig. 5. Mean F_1 standard deviation for the end-to-end VLM ensemble (solid) vs. raw VLM outputs (dotted), showing top-performing (Flux.2, Qwen 2511) and the least stable (Qwen) models.

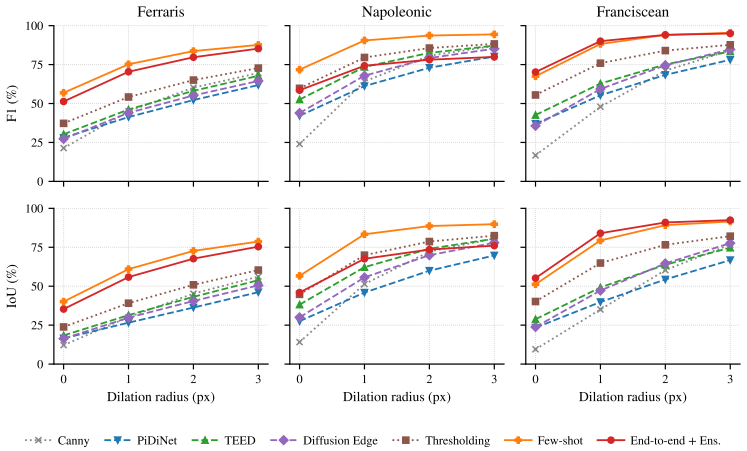


Fig. 6. Mean F_1 (top) and mean IoU (bottom) for increasing ground truth dilation radii $r \in 0, .., 3$ across all baselines and models.

Table 4. Few-shot approach by Sterzinger et al. [22] compared to the end-to-end ensemble pipeline across radii 0–3 across the three datasets.

	r0		**r1**		**r2**		**r3**	
Model	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU
Ferraris Map								
Few-shot	**56.87**	**40.09**	**75.21**	**60.96**	**83.66**	**72.66**	**87.64**	78.67
Flux.1	50.65	34.67	69.43	54.57	78.83	66.35	84.64	74.43
Flux.2	54.23	38.10	73.64	59.95	82.35	71.69	87.24	**78.92**
Flux.2 klein	50.01	34.04	69.61	54.71	79.51	67.20	85.20	75.09
Qwen	46.34	30.97	64.80	49.44	74.97	61.34	81.81	70.24
Qwen 2511	55.24	38.84	74.53	60.60	82.91	71.83	87.36	78.31
Napoleonic Land Registry								
Few-shot	**71.75**	**56.64**	**90.50**	**83.35**	**93.67**	**88.64**	94.39	89.84
Flux.1	66.03	50.86	85.13	76.95	89.19	83.21	91.10	86.09
Flux.2	66.47	50.75	87.17	78.38	93.37	88.44	**95.22**	**91.62**
Flux.2 klein	59.47	43.16	81.38	69.75	89.92	82.52	93.47	88.32
Qwen	45.99	33.68	60.60	50.10	65.44	55.14	68.75	58.40
Qwen 2511	60.84	45.45	83.58	74.53	90.21	84.12	92.79	87.85
Franciscean Cadastre								
Few-shot	67.50	51.26	88.23	79.39	94.06	89.20	95.41	91.58
Flux.1	70.31	55.51	89.42	83.46	92.97	89.61	93.77	91.01
Flux.2	75.03	60.23	**94.99**	**90.63**	**98.31**	**96.76**	**98.87**	**97.83**
Flux.2 klein	70.86	55.09	92.39	86.03	97.36	94.95	98.46	97.02
Qwen	55.84	41.70	72.96	63.23	77.72	69.49	79.74	71.91
Qwen 2511	**75.45**	**60.93**	94.57	90.25	97.49	95.60	97.94	96.44

6.1 VLM Fine-tuning and Domain Adaption

Finally, we evaluate the impact of LoRA-based fine-tuning on end-to-end segmentation. We contrast the fine-tuned models against two zero-shot configurations: the raw VLM output and the ensemble utilizing a majority vote mechanism ($K \geq 6$). By reporting mean F_1 and IoU, we analyse how task-specific fine-tuning influences both the raw generative fidelity and the spatial consistency of the segmentation masks.

Table 5. Mean LoRA performance (r=0) on the the Franciscean corpus.

model	F_1	Std	Best	Worst	IoU	Std	Best	Worst
Flux.2	68.89	7.12	81.28	35.82	52.96	7.76	68.47	21.82
Flux.2 LoRA	70.42	4.69	79.23	57.80	54.55	**5.48**	65.60	40.65
Flux.2 + Ens.	**75.03**	**4.61**	**82.44**	**63.95**	**60.23**	5.76	**70.12**	**47.00**
Flux.2 LoRA + Ens.	72.42	5.20	80.26	61.81	57.01	6.31	67.02	44.73
Qwen 2511	59.93	16.70	81.39	0.00	44.54	14.95	68.62	0.00
Qwen 2511 LoRA	66.23	6.29	77.80	50.23	49.83	**6.95**	63.67	33.54
Qwen 2511 + Ens.	**75.45**	**6.16**	**82.84**	**58.68**	**60.93**	7.46	**70.71**	**41.53**
Qwen 2511 LoRA + Ens.	70.97	6.61	80.82	57.22	55.39	7.89	67.81	40.08

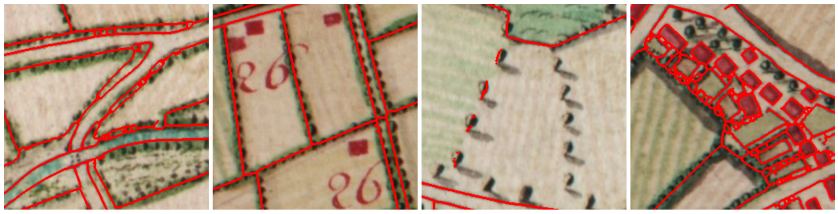
7 Discussion

The evaluation in Table 1 highlights the impact of VLM-driven canonical synthesis in lowering the algorithmic barrier for cartographic extraction. By standardizing archival noise, this approach allows simple intensity thresholding to outperform sophisticated detectors like TEED and PiDiNet. This suggests that VLM-based pre-processing can simplify the extraction tasks to a point where specialized feature detection becomes secondary to raw semantic cleaning.

However, this two-stage pipeline is consistently surpassed by end-to-end ensemble models. High-capacity architectures, specifically Qwen 2511 and Flux.2, prove most effective, with the superior performance of Flux.1 over its smaller counterpart Flux.2 [klein] reinforcing the importance of model scale. The strength of this approach lies in the ensemble mechanism, which filters the stochastic fluctuations of individual generative passes (see Figure 3). This is most evident in Qwen 2511's performance on the Franciscean corpus, where a 16.39 point IoU gain over the raw mean demonstrates that voting can overcome isolated generative errors (see Table 3).

The Ferraris corpus presents the most significant challenge for both baselines and end-to-end approaches due to its high level of illustrative detail. While dense vegetation markers are reliably synthesized into continuous boundaries, systematic errors persist: (i) morphological merging of proximal parallel lines (e.g. road and tree borders), (ii) omission of structural footprints, (iii) failure to resolve sparse vegetation patterns, and (iv) symbology interference, where building edges and dotted boundaries bleed into surrounding vegetation markers (see Figure 7). These errors suggest that while VLMs show promise in topological inference, they remain susceptible to the stylistic noise inherent in non-standardized archival cartography. See Figures B.8, B.9, and B.10 for examples of errors in the other corpora.

A distinct spatial trade-off emerges when comparing VLM-driven line segmentation to the few-shot baseline. While the baseline maintains a lead on the Napoleonic corpus at strict radii (r = 0–2), the performance gap narrows and eventually flips at r = 3 (see Table 4), indicating a shift in priority from pixel-perfect matching to topological continuity. On the more complex Franciscean dataset, Qwen 2511 (r = 0) and Flux.2 (r = 1–3) perform significantly bet-



(a) (b) (c) (d)

Fig. 7. End-to-end ensemble errors in the Ferraris corpus: (a) feature merging, (b) structural omission, (c) sparse vegetation failure, (d) textural boundary bleed.

ter, suggesting that generative models maintain line integrity even when local discriminative features are ambiguous.

Finally, in the context of our Franciscan cadastre dataset, the comparison between LoRA-adapted models and the ensemble method reveals a trade-off between reliability and accuracy (see Table 5). LoRA-adaption exhibits the lowest standard deviation and significantly improves worst-case outcomes compared to their base counterparts. However, despite surpassing the base models, they appear prone to systematic bias where consistent errors prevent higher scores. In contrast, while zero-shot base models are highly stochastic, as evident in their high variance and occasional catastrophic failures, their errors are non-systematic. Consequently, the majority vote ensemble effectively filters this generative noise (see Figure 3) to achieve the highest overall F_1 and IoU scores.

8 Conclusion

This paper demonstrates the effectiveness of VLM-driven models as unified engines for historical cartographic extraction. Our results indicate that while VLM-driven canonical synthesis effectively lowers the algorithmic barrier for line segmentation, the highest performance is achieved through an end-to-end ensemble approach. By leveraging majority vote aggregation, we successfully mitigate generative stochasticity and thus address an inherent problem of generative models. Furthermore, our findings suggest that while generative models may lack pixel-perfect precision, they offer distinct advantages in capturing topological structures compared to discriminative baselines, particularly on highly complex corpora like the Franciscean Cadastre. While task-specific fine-tuning via LoRA offers a path toward increased reliability, the zero-shot ensemble mechanism remains the most effective strategy for maximizing extraction accuracy in our experiments. In summary, VLM-driven synthesis offers a scalable, annotation-free approach for transforming historical cartography into machine-readable data, significantly reducing the need for labour-intensive manual labelling.

Disclosure of Interests. The authors have no competing interests to declare.

Acknowledgments. This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1763723 and 10.55776/FG3100.

References

1. Ares Oliveira, S., di Lenardo, I., Tourenc, B., Kaplan, F.: A deep learning approach to Cadastral Computing. In: Digital Humanities Conference. Utrecht, Netherlands (Jul 2019), https://hal.science/hal-03988983
2. Black Forest Labs: Flux.2: Advancing rectified flow matching for high-resolution synthesis. Tech. rep., Black Forest Labs (2025), https://blackforestlabs.ai/flux-2-technical-report/
3. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851
4. Chiang, Y.Y., Duan, W., Leyk, S., Uhl, J.H., Knoblock, C.A.: Training Deep Learning Models for Geographic Feature Recognition from Historical Maps, pp. 65–98. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-319-66908-3_4, https://doi.org/10.1007/978-3-319-66908-3_4
5. Dhar, D.B., Chanda, B.: Extraction and recognition of geographical features from paper maps. International Journal of Document Analysis and Recognition (IJDAR) **8**(4), 232–245 (2006). https://doi.org/10.1007/s10032-005-0010-9, https://doi.org/10.1007/s10032-005-0010-9
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://arxiv.org/abs/2403.03206
7. Freeman, H., Pieroni, G.G.: Map Data Processing: Proceedings of a NATO Advanced Study Institute on Map Data Processing Held in Maratea, Italy, June 18–29, 1979. Academic Press, New York (1980). https://doi.org/10.1016/C2013-0-10688-5
8. Heitzler, M., Hurni, L.: Cartographic reconstruction of building footprints from historical maps: A study on the swiss siegfried map. Transactions in GIS **24**, 442–461 (01 2020). https://doi.org/10.1111/tgis.12610
9. Henderson, T.C., Linton, T.: Raster map image analysis. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 376–380 (2009). https://doi.org/10.1109/ICDAR.2009.31
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
12. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
13. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
14. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation (2024), https://arxiv.org/abs/2402.09353

15. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022), https://arxiv.org/abs/2209.03003
16. Miyoshi, T., Li, W., Kaneda, K., Yamashita, H., Nakamae, E.: Automatic extraction of buildings utilizing geometric features of a scanned topographic map. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. vol. 3, pp. 626–629 Vol.3 (2004). https://doi.org/10.1109/ICPR.2004.1334607
17. Petitpierre, R., Kaplan, F., di Lenardo, I.: Generic semantic segmentation of historical maps. In: Ehrmann, M., Karsdorp, F., Wevers, M., Andrews, T.L., Burghardt, M., Kestemont, M., Manjavacas, E., Piotrowski, M., van Zundert, J. (eds.) Proceedings of the Conference on Computational Humanities Research, CHR2021, Amsterdam, The Netherlands, November 17-19, 2021. CEUR Workshop Proceedings, vol. 2989, pp. 228–248. CEUR-WS.org (2021), https://ceur-ws.org/Vol-2989/long_paper27.pdf
18. Pivac, D., Roić, M., Križanović, J., Paar, R.: Availability of historical cadastral data. Land **10**, 917 (08 2021). https://doi.org/10.3390/land10090917
19. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), https://arxiv.org/abs/2408.00714
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
21. Soria, X., Li, Y., Rouhani, M., Sappa, A.D.: Tiny and efficient model for the edge detection generalization (2023), https://arxiv.org/abs/2308.06468
22. Sterzinger, R., Peer, M., Sablatnig, R.: Few-Shot Segmentation of Historical Maps via Linear Probing of Vision Foundation Models, p. 425–442. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-04624-6_25, http://dx.doi.org/10.1007/978-3-032-04624-6_25
23. Su, Z., Liu, W., Yu, Z., Hu, D., Liao, Q., Tian, Q., Pietikäinen, M., Liu, L.: Pixel difference networks for efficient edge detection (2021), https://arxiv.org/abs/2108.07009
24. Vervust, Soetkin: Deconstructing the Ferraris maps (1770-1778) : a study of the map production process and its implications for geometric accuracy. Ph.D. thesis, Ghent University (2016)
25. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
26. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
27. Xia, X., Balestriero, R., Zhang, T., Zhou, Y., Ding, A., Saini, D., Hurni, L.: Mapsam2: Adapting sam2 for automatic segmentation of historical map images and time series (2025), https://arxiv.org/abs/2510.27547
28. Xia, X., Zhang, D., Song, W., Huang, W., Hurni, L.: Mapsam: Adapting segment anything model for automated feature detection in historical maps (2024), https://arxiv.org/abs/2411.06971
29. Ye, Y., Xu, K., Huang, Y., Yi, R., Cai, Z.: Diffusionedge: Diffusion probabilistic model for crisp edge detection (2024), https://arxiv.org/abs/2401.02032

30. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization (2015), https://arxiv.org/abs/1506.06579

A Tables

Table A.6. Standard deviation of F_1 and IoU at ground truth dilation radius $r = 0$ across all VLM generated canonical images for the three map corpora.

	Canny		**PiDiNet**		**TEED**		**Diff. Edge**		**Thresh.**	
Model	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU	F_1	IoU
Ferraris Map										
Flux.1	6.51	3.98	8.11	5.25	8.75	5.87	8.53	5.71	10.27	7.03
Flux.2	6.26	3.92	6.35	4.25	7.81	5.63	**7.07**	**4.91**	10.27	8.45
Flux.2 klein	5.32	3.28	5.50	3.54	**6.96**	**4.79**	7.54	5.17	**8.60**	**6.41**
Qwen	**3.35**	**2.04**	**5.07**	**3.42**	12.45	9.27	11.55	8.39	10.70	9.49
Qwen 2511	4.84	3.01	5.77	3.75	9.96	7.20	8.06	5.63	8.89	6.82
Napoleonic Land Registry										
Flux.1	**6.29**	4.29	8.49	6.65	11.85	10.58	10.23	8.31	12.70	11.33
Flux.2	6.38	4.07	**6.80**	5.68	10.60	9.54	**10.01**	**7.96**	9.84	10.78
Flux.2 klein	6.49	**3.88**	7.35	**5.27**	**9.76**	**7.33**	12.76	9.05	**8.42**	**8.51**
Qwen	7.18	4.87	10.47	7.98	16.77	14.03	13.77	10.53	17.83	14.81
Qwen 2511	9.34	6.66	12.49	11.00	17.26	17.14	14.05	12.41	12.45	12.46
Franciscean Cadastre										
Flux.1	6.38	4.14	8.04	6.16	10.70	9.43	9.04	7.76	10.67	9.27
Flux.2	**2.63**	**1.43**	6.46	**4.17**	**6.91**	**4.46**	**6.60**	**3.88**	11.36	10.28
Flux.2 klein	4.30	2.48	**6.31**	4.74	7.09	5.49	8.85	6.09	**7.68**	**8.04**
Qwen	6.19	4.14	8.32	6.64	11.22	10.42	9.07	7.88	15.36	13.94
Qwen 2511	8.68	5.18	13.12	9.06	16.51	12.25	16.04	11.36	15.28	13.23

Table A.7. Standard deviation of F_1 and IoU as well as best and worst achieved scores at ground truth dilation radius $r = 0$ across all VLM generated segmentation images without ensemble for the three map corpora.

Model	F_1	Std.	Best	Worst	IoU	Std.	Best	Worst
Ferraris Map								
Flux.1	44.90	13.38	73.40	**2.55**	29.86	10.73	57.98	**1.29**
Flux.2	**47.27**	10.37	68.34	21.67	**31.55**	8.94	51.91	12.15
Flux.2 klein	46.57	**9.84**	65.26	17.74	30.88	**8.20**	48.44	9.74
Qwen 2511	40.48	14.90	**77.99**	4.09	26.54	12.54	**63.92**	2.09
Qwen	30.12	17.62	74.91	3.43	19.09	13.28	59.88	1.74
Napoleonic Land Registry								
Flux.1	62.78	13.55	79.99	0.66	46.99	12.74	66.65	0.33
Flux.2	**72.86**	**7.48**	**87.23**	39.59	**57.83**	8.92	**77.36**	24.68
Flux.2 klein	65.97	9.76	84.37	35.94	49.99	10.66	72.97	21.90
Qwen 2511	50.13	19.90	82.26	**0.00**	35.67	17.00	69.87	**0.00**
Qwen	11.59	12.38	62.93	1.98	6.70	**8.40**	45.92	1.00
Franciscean Cadastre								
Flux.1	60.10	16.66	76.48	0.62	44.62	14.18	61.92	0.31
Flux.2	**68.89**	**7.12**	81.28	35.82	**52.96**	**7.76**	68.47	21.82
Flux.2 klein	60.15	8.46	75.41	37.93	43.52	8.58	60.53	23.40
Qwen 2511	59.93	16.70	**81.39**	**0.00**	44.54	14.95	**68.62**	**0.00**
Qwen	38.98	25.16	78.52	2.31	27.30	19.98	64.63	1.17

B Figures

Fig. B.8. Qwen 2511 segmentation errors (r=0) on Ferraris maps: (a) persistent tree lines, (b) building omission, (c) persistent dotted lines, (d) feature merging.

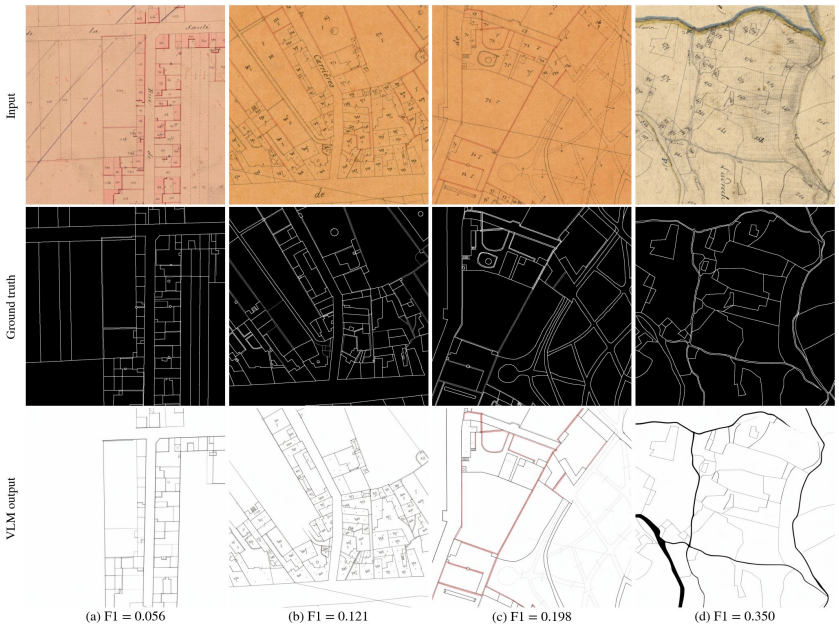


Fig. B.9. Qwen 2511 segmentation errors (r=0) on Napoleonic maps: (a) line omission, (b) persistent textual annotations, (c) persistent colours, (d) artefactual area-filling.

Fig. B.10. Qwen 2511 segmentation errors (r=0) on Franciscean maps: (a, b) hallucination, (c) sub-threshold lines, (d) persistent colours and false edges.

C Prompts

Listing C.1. An example prompt for Qwen 2511 canonical image synthesis on the Franciscean Cadastre, utilizing detailed task description and negative constraints to guide the process.

```
"A high-resolution, digitized restoration of the historical
cadastral map tile. The output must be a "clean" version, completely
stripped of all non-geometric data. Remove all handwritten and
printed text, including parcel numbers, letters, and margin
annotations. Delete all point-based symbols, such as tree icons,
vegetation markers, cross-hatched soil symbols, and terrain
textures. Remove all paper stains, foxing, and age-related
discolorations.
Retain the original colors of the background. All black parcel
boundary lines must be rendered as continuous, unbroken, and solid
paths. These lines should be uniformly thickened and reinforced to a
heavy, high-contrast weight to ensure optimal edge detection for GIS
processing. The final image should consist only of uniform color
blocks (inferred from the original background color) and bold,
simplified geometric borders."
```

Listing C.2. An example prompt for Flux.1 end-to-end segmentation on the Napoleonic corpus, utilizing an imperative, command-driven structure.

```
"Convert this historical cadastral map into a clean, high-contrast
binary line drawing. Isolate and keep only the dark, structural
parcel boundary lines. Complete broken lines, filling in small gaps.
Turn the background into pure flat white (#FFFFFF) and the boundary
lines into pure flat black (#000000).

Strictly remove the following:
    All handwritten text, numbers, and labels.
    All paper texture, discoloration, stains, and noise.
    All decorative symbols, trees, shrubbery.

The output should look like a digital architectural vector plan or a
binarized skeleton map suitable for edge detection with no symbols
and no text."
```