Price Prediction for MIT's WhatsApp Marketplace

15.776: Intensive Hands-on Deep Learning

Eric Christenson, Cesar Dori, Maria Angel Lobon, Franco Martino

December 2025

Abstract

We study multimodal product price prediction in the context of an informal, peer-to-peer marketplace: the MIT Sloan Buy & Sell WhatsApp group. In this group chat, students regularly post items for sale using a single image and a short, noisy text description, often consisting of abbreviations, informal language, or incomplete product information. Motivated by our own experience using this marketplace—where sellers frequently struggle to determine a reasonable listing price—we aim to build a model that assists users by suggesting a fair reference price directly from the image and text of a listing.

We formulate this task as a regression problem that leverages both visual and textual inputs. Images are encoded using a Vision Transformer (ViT) to capture cues such as product type, condition, and brand, while listing titles are encoded with BERT to extract semantic information from unstructured text. The resulting embeddings are concatenated and passed through a fully connected prediction head to produce a price suggestion.

We construct and curate a custom multimodal dataset from the MIT Sloan Buy & Sell WhatsApp group, consisting of real-world listings with paired images, text descriptions, and observed prices. Because this dataset is relatively small, we first leverage a large-scale multimodal Amazon product dataset from Hugging Face to learn robust representations and stabilize training. We then transfer these representations to the WhatsApp marketplace domain via fine-tuning on our curated dataset. Our results show that combining image and text information significantly improves price prediction accuracy over unimodal baselines, highlighting the value of multimodal learning for practical price suggestion tools in low-data, real-world marketplace settings.

1 Problem Statement

Every year, a large number of MIT students join and leave Cambridge, generating a highly active secondhand market for furniture, electronics, and household items. Most of these transactions take place in informal WhatsApp groups such as the MIT Sloan Buy & Sell chat, where sellers typically post a single image and a short, noisy text description of the item.

For new members of the group, setting a reasonable price is often challenging. Sellers frequently lack reference points, are unfamiliar with typical resale prices, or are unsure how to account for factors such as condition, brand, and visual quality.

In this project, we study whether multimodal learning can help address this problem by recommending a reference price using only the information available in a typical WhatsApp listing: an image and a short text description. By jointly leveraging visual and textual cues, we aim to approximate how humans assess value in informal peer-to-peer resale settings.

2 Data

This project relies on two complementary datasets: (1) a small, domain-specific dataset constructed from MIT WhatsApp resale conversations, and (2) a large-scale external multimodal dataset from Hugging Face used for representation learning and pretraining.

2.1 MIT WhatsApp Buy & Sell Dataset

Our primary in-domain dataset is derived from listings posted in the MIT Sloan Buy & Sell WhatsApp group, an informal peer-to-peer marketplace used by students to resell items such as furniture, electronics, and household goods. Each listing typically consists of a single product image, a short and noisy textual description, and an asking price. The text is often informal, abbreviated, or incomplete, reflecting the conversational nature of WhatsApp messages.

Because the data originate from unstructured chat conversations, substantial preprocessing is required to construct a usable dataset. While this dataset accurately reflects the real-world deployment setting, its relatively small size poses challenges for training high-capacity deep learning models from scratch.

2.1.1 Dataset Construction Pipeline

The main challenge in building the dataset is that WhatsApp data is highly unstructured: free-form text, inconsistent price formats, multiple images per post, and links to external documents. We designed a semi-automated multimodal extraction pipeline using GPT-4o-mini to convert this noisy data into a clean structured dataset.

WhatsApp Extraction The "Sloan Buy / Sell 26+ 25s" group chat was exported, containing messages, timestamps, images, and links. We use a two-pass extraction strategy:

Pass 1: Text-only extraction The model analyzes message text to:

- detect when a user is selling an item,
- extract a clean description (removing price tokens),
- normalize price expressions ($50, "50 bucks", "1.5k", etc.),

- propose associated images based on temporal proximity and sender.

Pass 2: Multimodal validation For each candidate item and its proposed images, the model visually checks whether the image corresponds to the extracted description. If none match, the item is retained without an image. Manual spot checks were performed to calibrate prompt quality and extraction consistency.

PDF and Google Slides Extraction Some sellers share Google Slides documents containing multiple products per page. These documents are converted to PDF, and text and images are extracted page by page. A similar two-pass LLM approach identifies product blocks and validates image–text consistency.

Final Dataset Assembly Items extracted from WhatsApp messages and PDF documents are merged into a single dataset. We remove entries missing valid text or price information, deduplicate repeated listings, and perform a final multimodal consistency check. The resulting dataset is stored as a structured CSV file with fields `description`, `price`, and `image_ref`.

Dataset Summary The final dataset contains 174 unique items, of which 65 (37%) include an associated image. Prices range from \$5 to \$3100, with a median price of \$47.50.

2.2 External Amazon Multimodal Dataset

To address the limited size of the WhatsApp dataset, we leverage the `amazon-product-descriptions-vlm` dataset from Hugging Face. This dataset contains Amazon product listings with paired images, descriptions, and prices across a wide range of categories.

Although Amazon listings differ from WhatsApp posts in tone, structure, and pricing dynamics, they share the same fundamental multimodal structure. We use this dataset for representation learning before fine-tuning on the WhatsApp domain.

3 Approach

Our objective is to build a multimodal deep learning model that predicts a fair resale price from an item's image and textual description. Because the WhatsApp dataset is small and noisy, we adopt a two-stage strategy combining large-scale pretraining with domain adaptation.

Throughout, we predict log(price) rather than raw price to stabilize optimization and reduce the influence of extreme values.

3.1 Stage 1: Multimodal Representation Learning

Text descriptions are encoded using a pretrained transformer-based language model (BERT), while images are encoded using a Vision Transformer (ViT). The resulting embeddings are concatenated and passed to a multilayer perceptron regression head that outputs a scalar log-price prediction.

3.2 Stage 2: Pretraining on the Amazon Dataset

The model is first trained on the Amazon dataset using supervised regression with frozen encoders and a trainable regression head. This forces the model to learn how to combine pretrained representations for price estimation.

3.3 Stage 3: Domain Adaptation on WhatsApp Data

We fine-tune the pretrained model on the WhatsApp dataset using a lower learning rate. Depending on the setting, encoders are either kept frozen or partially unfrozen to allow limited domain-specific adaptation.

4 Results

We evaluate the proposed model under two complementary settings that reflect different stages of the training pipeline and serve distinct purposes. The **global test set** corresponds to a held-out split of the Amazon multimodal dataset and evaluates the model after pretraining, before any fine-tuning on WhatsApp data. The **local test set** corresponds to a held-out split of the MIT WhatsApp dataset and evaluates the model after fine-tuning on in-domain data. This local evaluation is the primary setting of interest, as it reflects the real deployment scenario.

Table 1: Error Metrics on Global (Amazon) and Local (WhatsApp) Test Sets

Test Set	Epochs	RMSE	MAE (Log)
Global (Amazon)	10	0.88	0.71
Local (WhatsApp)	10	1.39	1.16

Table 4 summarizes the error metrics for both evaluation settings. Overall, the model achieves strong predictive performance on the global test set and reasonable performance on the local test set, given the substantially smaller size and higher noise level of the WhatsApp data.

Global Evaluation (Amazon Dataset). We first examine performance on the global test set, which provides a clean, large-scale benchmark for multimodal price prediction. As shown in Figures 1 and 3, prices in this dataset span a wide range but are relatively well distributed. Figure 2 summarizes model performance, with predicted prices closely tracking ground-truth values and residuals centered near zero. The low RMSE and MAE indicate that the pretrained model learns meaningful relationships between visual appearance, textual descriptions, and product prices. Qualitative examples in Figure 4 further confirm that predictions are consistent and reasonable across diverse product categories. We also observe that increasing the number of training epochs from 3 to 10 improves performance, suggesting that the regression head benefits from additional optimization once representations are fixed.

Local Evaluation (WhatsApp Dataset). We now turn to the local test set, which reflects the target deployment environment. As illustrated in Figure 5, log-prices in the WhatsApp dataset are less evenly distributed, reflecting the small dataset size and the heterogeneity of items, image quality, and listing descriptions. Correspondingly, Figure 6 shows increased dispersion in residuals compared to the global setting. While fine-tuning enables the model to adapt to resale-specific price levels and patterns, prediction errors are higher than in the global evaluation. This gap highlights the intrinsic difficulty of pricing in informal peer-to-peer markets, where data are sparse and important pricing factors are not fully observable. Nevertheless, the model continues to capture general pricing trends, supporting its use as a reference price suggestion tool rather than an exact valuation mechanism.

5 Lessons Learned

5.1 What Worked

- **Data quality mattered more than model complexity.** In a noisy WhatsApp setting, improvements from cleaning, deduplication, and multimodal consistency checks were essential to make training stable.
- **Transfer learning was necessary.** Pretraining on a large external product dataset provided a strong initialization and made learning feasible with limited in-domain data.
- **Multimodality helped in practice.** Images and short text descriptions carried complementary signals (e.g., condition/quality vs. item identity)

5.2 What Was Hard

- **Heterogeneity limits accuracy.** A single global model must learn pricing across very different categories and conditions, which increases variance and makes rare items difficult to price.
- **Unobserved factors remain.** Negotiation dynamics, urgency, and idiosyncratic buyer preferences are not captured by image/text, so predictions should be interpreted as reference prices.

5.3 Extensions

- **Category-aware modeling:** specialize models by product type (e.g., furniture vs. electronics) or add an inferred category as an input.
- **Add lightweight context:** incorporate signals available in chat data (e.g., posting time, historical price ranges within the group) to reduce ambiguity.
- **Scale the pipeline:** apply the same dataset construction framework to additional university resale groups to evaluate generalization across communities.

Appendix

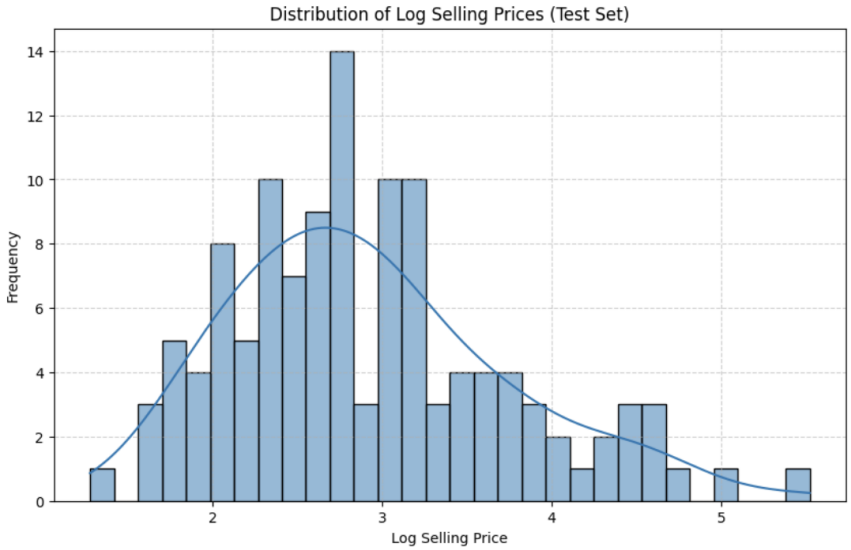

Figure 1: Global log prices of the test set

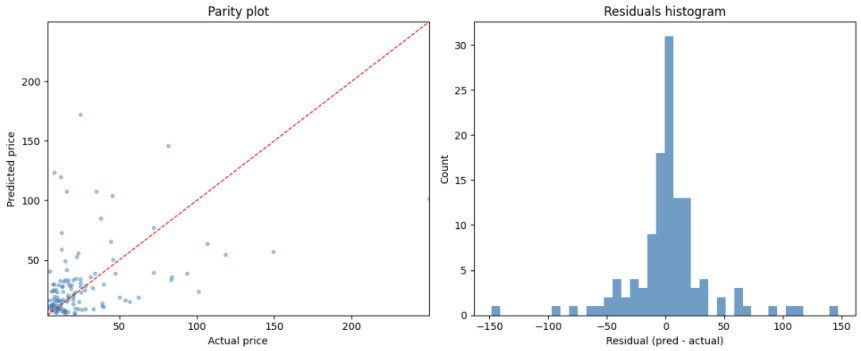

Figure 2: Global parity and residuals test set

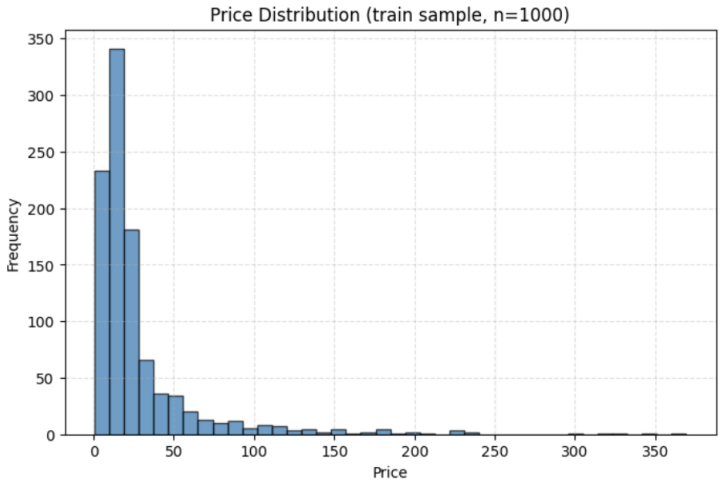

Figure 3: Global price distribution

Example 1
Description: Kurio Glow Smartwatch for Kids with Bluetooth, Apps, Camera & Games, Blue (ID: 002e4642d3ead5ecdc9958ce0b3a5a79)
Pred price: 36.53747205311434
Actual price: 31.3

Example 2
Description: Qualatex Foil Balloon 13955 Little Man Bow-tie, 18", Multicolor (ID: 022613ca2afc8d2779b77fba98ab4a46)
Pred price: 7.754531468328155
Actual price: 6.01

Example 3
Description: Disney "Cars 2" Paint By Numbers, Party Favor (ID: 04b9ad587c23b5bb58cbbefab1c64d34)
Pred price: 12.587846329994422
Actual price: 7.39

Figure 4: Example items from global test set and predictions

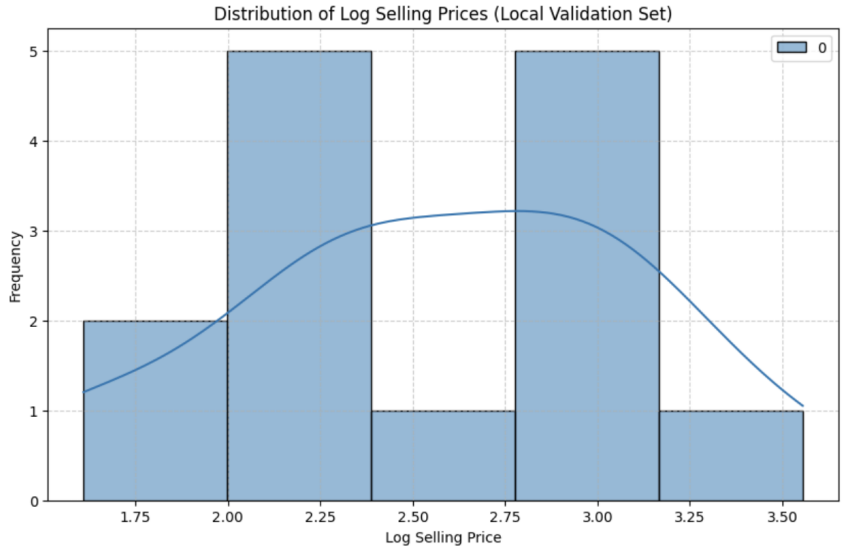

Figure 5: Local log prices test set

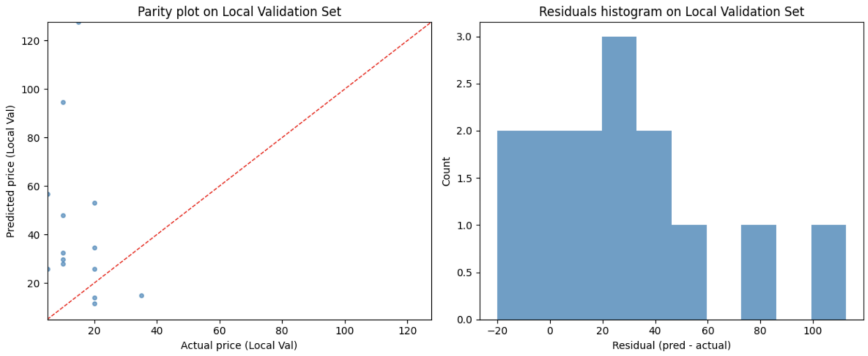

Figure 6: Local parity and residuals test set

6 Reproducibility

- **External dataset:** `https://huggingface.co/datasets/philschmid/amazon-product-descriptions-vlm`.
- **Code repository:** `https://github.com/mariangellobon/Multimodal-Transfer-Learning-Price-Predict`
- **Google Colab notebook:** `https://colab.research.google.com/drive/1VrUFsL36K370xIQIYItgPLF42m6f?authuser=3#scrollTo=u0l08YpueD0e`.

The Google Colab notebook contains the full modeling and training pipeline. Some steps in the notebook rely on local auxiliary files (e.g., intermediate artifacts used during WhatsApp dataset construction). These files are available or can be regenerated using the scripts provided in the GitHub repository, ensuring that all results in this project can be reproduced end-to-end.