TraveLLaMA: A Multimodal Travel Assistant with Large-Scale Dataset and Structured Reasoning

Meng Chu[1], **Yukang Chen**[2], **Haokun Gui**[1], **Shaozuo Yu**[2], **Yi Wang**[3], **Jiaya Jia**[1]

[1]Hong Kong University of Science and Technology
[2]Chinese University of Hong Kong
[3]Shanghai AI Laboratory

Abstract

Tourism and travel planning increasingly rely on digital assistance, yet existing multimodal AI systems often lack specialized knowledge and contextual understanding of urban environments. We present **TraveLLaMA**, a specialized multimodal language model designed for comprehensive travel assistance. Our work addresses the fundamental challenge of developing practical AI travel assistants through three key contributions: (1) TravelQA, a novel dataset of 265k question-answer pairs combining 160k text QA from authentic travel sources, 100k vision-language QA featuring maps and location imagery, and 5k expert-annotated Chain-of-Thought reasoning examples; (2) Travel-CoT, a structured reasoning framework that decomposes travel queries into spatial, temporal, and practical dimensions, improving answer accuracy by 10.8% while providing interpretable decision paths; and (3) an interactive agent system validated through extensive user studies. Through fine-tuning experiments on state-of-the-art vision-language models (LLaVA, Qwen-VL, Shikra), we achieve 6.2-9.4% base improvements, further enhanced by Travel-CoT reasoning. Our model demonstrates superior capabilities in contextual travel recommendations, map interpretation, and scene understanding while providing practical information such as operating hours and cultural insights. User studies with 500 participants show TraveLLaMA achieves a System Usability Scale score of 82.5, significantly outperforming general-purpose models and establishing new standards for multimodal travel assistance systems.

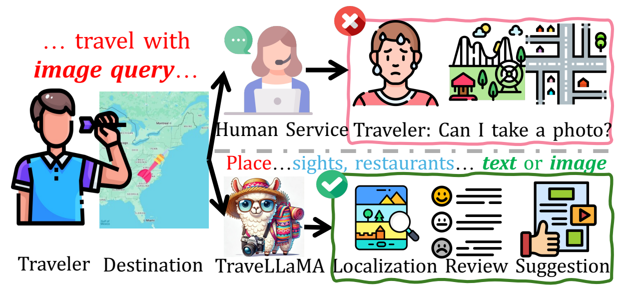

Figure 1: TraveLLaMA, an advanced multimodal AI travel assistant that seamlessly processes both text and image-based queries. This powerful system enables travelers to plan trips efficiently by providing contextual responses including human service information, localization details, and personalized recommendations based on visual inputs and textual questions about destinations, sights, and restaurants.

Introduction

Travel planning exemplifies the complexity of real-world AI applications, requiring simultaneous understanding of visual scenes, geographical contexts, and practical constraints. While large language models (LLMs) have achieved remarkable success in many domains (Touvron et al. 2023; Zhao et al. 2023b), they struggle with travel assistance due to a critical gap: the absence of multimodal datasets that capture the inherently visual and contextual nature of travel planning (Schumann et al. 2024; Li et al. 2024).

This limitation is particularly acute because effective travel assistance demands integration across multiple modalities. Recommending a restaurant, for instance, requires understanding its location on a map, recognizing ambiance from photos, interpreting user reviews, and considering operational constraints—a level of multimodal reasoning that current datasets fail to support (Yang et al. 2024; Vivanco Cepeda, Nayak, and Shah 2023). Similarly, planning a day-long itinerary involves analyzing distances between attractions, understanding transportation options from visual maps, recognizing architectural styles from images, and incorporating temporal constraints like opening hours and peak times. General-purpose models, despite their broad training, lack the domain-specific knowledge to connect visual landmarks with practical travel information (Li et al. 2023b; Maheshwary, Paul, and Sohoney 2024).

Furthermore, the challenge extends beyond simple multimodal understanding to requiring culturally-aware and contextually-appropriate reasoning. A temple visit in Kyoto demands different preparations than a beach day in Bali, yet current AI systems often provide generic advice that fails to capture these nuances. The absence of structured reasoning frameworks means that even when models can process individual modalities, they struggle to synthesize information coherently—leading to recommendations that may be factually correct but practically infeasible, such as suggesting attractions without considering accessibility for elderly travelers or recommending outdoor activities without accounting for seasonal weather patterns.

Figure 2: The TravelQA dataset features iconic landmarks and destinations across major cities worldwide, connecting locations like San Francisco, New York, Paris, Rome, Berlin, Bangkok, Singapore, Shanghai, and Beijing through a global travel network.

While structured travel datasets remain scarce, the web contains abundant unstructured content across forums, review sites, and mapping services. We leverage modern LLMs to transform this fragmented information into high-quality multimodal training data (Xie et al. 2023a), enabling cost-effective dataset creation at scale.

To address these challenges, we present TraveLLaMA (Figure 1), a specialized multimodal system for AI-powered travel assistance. Our contributions are:

- **TravelQA Dataset.** We create the first large-scale multimodal travel dataset with 265k QA pairs: 160k text-based pairs from travel forums, 100k vision-language pairs with maps and photos, and 5k expert-annotated CoT reasoning examples.
- **Travel-CoT Reasoning.** Beyond base improvements of 6.2-9.4%, our Travel-CoT framework decomposes queries into spatial, temporal, and practical dimensions, achieving 10.8% accuracy gain with interpretable reasoning paths.
- **Interactive Agent System.** Our ReAct-based agent integrates real-time services for dynamic planning, validated by 500 users with a SUS score of 82.5, demonstrating superior usability for complex travel tasks.

Related Work

Large Language Models. Large Language Models (LLMs) (Shafik 2024; Touvron et al. 2023; Zeng et al. 2022; Zhao et al. 2023b; Chu, Li, and Chua 2025) show impressive capabilities yet underperform in specialized domains, prompting development of domain-specific models for finance (Wu et al. 2023), medicine (Singhal et al. 2023), mathematics (Azerbayev et al. 2023), and geospatial applications (Deng et al. 2024; Wang et al. 2023). These models have been studied for spatial understanding (Gurnee and Tegmark 2023; Momennejad et al. 2023; Yamada et al. 2023) and geographic representation (Godey, de la Clergerie, and Sagot 2024; Manvi et al. 2023). LLM-powered language agents like AutoGPT (Yang, Yue, and He 2023), BabyAGI (Nakajima 2023), and HuggingGPT (Shen et al. 2023) decompose complex tasks through Memory, Tool-use, and Planning modules, utilizing memory summarization (Chen et al. 2023a; Zhou et al. 2023; Liang et al. 2023; Yu et al. 2025), retrieval techniques (Andreas 2022; Park et al. 2023; Zhong et al. 2024), and tool-augmentation (Nakano et al. 2021; Lu et al. 2023; Ge et al. 2023; Xie et al. 2023b). Our approach focuses on understanding physical urban spaces and spatial reasoning to solve real urban challenges, demonstrating the geospatial knowledge capabilities in pre-trained language models.

Vision-Language Models. The evolution of vision-language models has seen significant advancement, beginning with early works in visual-semantic embedding (Frome et al. 2013; Kiros et al. 2014; Chu et al. 2024) and progressing to more sophisticated architectures. Recent models like GPT-4V (OpenAI 2023) and LLaVA (Liu et al. 2023b) have

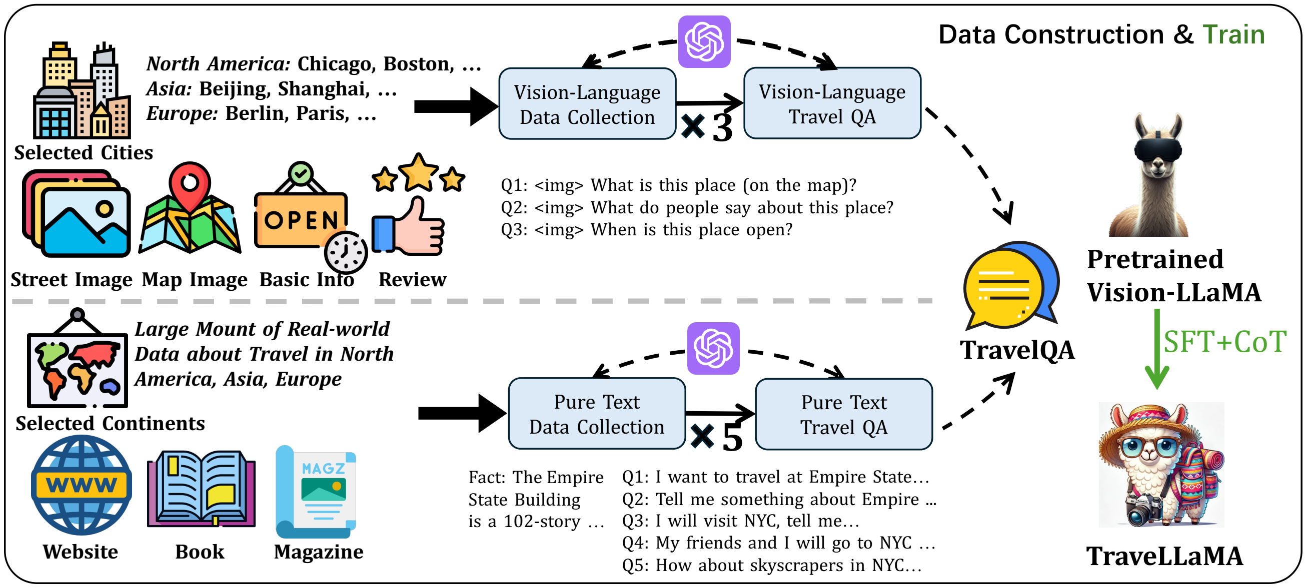

Figure 3: TraveLLaMA's data construction and training process combines vision-language and text-based travel QA from diverse global sources through multi-round collection and fine-tuning.

demonstrated unprecedented capabilities in joint text-vision understanding. However, these general-purpose models face significant challenges in travel-specific applications. Recent work has attempted to address these issues through specialized architectures (Yang et al. 2023) and domain-specific pre-training (Zhao et al. 2023a), but significant challenges remain in achieving human-level understanding of travel-related visual content (Brown et al. 2023).

LLMs in Urban Applications. Travel AI systems have evolved from basic route planning and POI recommendation to advanced systems like GeoLLM (Li et al. 2023b) and GeoReasoner (Li et al. 2024). Recent work focuses on multi-modal integration (Yang et al. 2024; Vivanco Cepeda, Nayak, and Shah 2023), though challenges persist in temporal reasoning and cross-modal alignment (Schumann et al. 2024). Specialized developments span cultural context, accessibility, and personalization (Maheshwary, Paul, and Sohoney 2024), while real-world implementations reveal challenges in uncertainty handling and system reliability. Applications extend to traffic safety analysis, geospace understanding, and accident reporting (Chu et al. 2024; Feng et al. 2024). While TravelPlanner (Xie et al. 2024) focuses on text-only travel planning, our system addresses multimodal single-day urban itinerary planning with multi-day extensibility, motivated by current limitations in handling complex travel planning scenarios (Xie et al. 2023a).

TravelQA

Dataset Overview

Figure 2 and Table 1 present TravelQA, a multimodal resource containing 265k QA pairs. To clarify points raised by reviewers, we explicitly describe the numerical composition: the dataset includes 160k text-only QA pairs, 100k vision–language QA pairs, and 5k expert-annotated CoT examples. The 160k text QA pairs come from 26k factual units expanded into five diverse questions each (130k), plus 30k augmented QAs focusing on practical constraints such as safety, cost, and accessibility. The 100k vision–language QA pairs are constructed from 20k POIs, each with an average of 4–5 street-view or map images; GPT-4 generates three question types *per image*, yielding roughly $20k \times 4{-}5 \times 3 \approx 100k$ pairs. The 5k CoT examples represent a single set where each item contains spatial, temporal, and practical reasoning, rather than 15k separate labels. To prevent leakage across modalities, **all splits are POI-disjoint**: every POI and its associated images and metadata appear in exactly one

Category	Subcategory	All	Train	Test
QA Format	Text	160k	128k	32k
	Vis-Lang	100k	80k	20k
	CoT	5k	5k	–
	Total	**265k**	**213k**	**52k***
Locations	Attractions	70k	56k	14k
	Dining	52k	42k	10k
	Living	39k	31k	8k
	Transportation	26k	21k	5k
	Cultural	39k	31k	8k
	Practical	34k	27k	7k
Visual Elements	Map	40k	32k	8k
	Street	60k	48k	12k
CoT Annotations	Spatial	5k	5k	–
	Temporal	5k	5k	–
	Practical	5k	5k	–

Table 1: Complete Dataset Composition and Usage Analysis

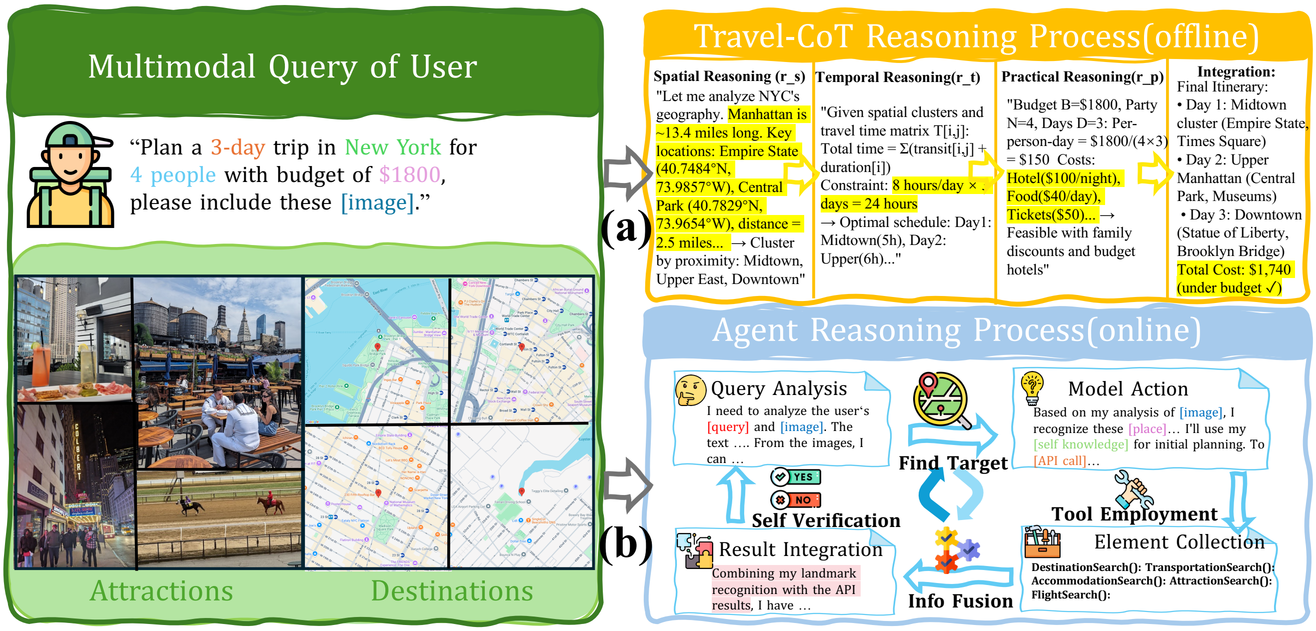

Figure 4: TraveLLaMA uses a reasoning and acting process to create travel plans. When a user submits a text-image query, the system could do the reasoning process offline or online, analyzing both components, identifies locations, employs specialized tools through API calls, and generates detailed itineraries with budget calculations matching user requirements.

split. The dataset spans 35+ cities across North America, Asia, and Europe, covering varied cultural and geographic contexts. Text-based answers average 45.6 words, while visual answers average 25–28 words, with all QA pairs passing multi-stage verification to ensure quality and factual consistency.

Dataset Construction Methodology

Pipeline Overview. As shown in Figure 3, TravelQA is built through parallel text and vision–language tracks processing street images, maps, descriptions, reviews, and structured facts. These supervised signals enable TraveLLaMA to answer a wide variety of travel queries.

Vision–Language QA Construction. For each of the 20k POIs, we collect 4–5 images, descriptive metadata, operational information, and filtered reviews. GPT-4 generates three QA types per image—identification, experience, and practical—resulting in approximately 100k QA pairs. Answers average 25–28 words and capture both visual content and contextual information.

Text-Based QA Generation. From 26k facts extracted from travel guides and online resources, GPT-4 produces five diverse questions per fact, alongside 30k additional QAs emphasizing practical constraints. Each QA pair undergoes a three-layer verification procedure combining rule checks, semantic consistency checks, and manual spot reviews, yielding 160k high-quality text QAs.

Travel-CoT Data Construction. To teach structured reasoning, we annotate 5,000 complex queries requiring multi-factor decision making. Travel experts provide explicit chains-of-thought containing *spatial*, *temporal*, and *practical* reasoning steps within each example, followed by final integrated answers. This unified annotation format ensures coherent multi-dimensional reasoning rather than separate labels.

Feature	TravelQA (Ours)	TravelPlanner
Modality	Text+Visual	Text-only
Query Scale	265k	1,225
Visual Content	Map, Street Photos	None
Category	6 structured categories (Attractions, Dining, etc.)	Unstructured

Table 2: Comparative Analysis of Datasets

Discussion. Table 2 highlights the breadth of TravelQA compared with prior resources. By integrating text, maps, street images, and structured CoT annotations, the dataset enables models to jointly reason over spatial layout, schedules, cultural context, and practical constraints—capabilities essential for real-world travel assistance systems.

Method

We develop TraveLLaMA by augmenting vision–language models with structured reasoning (Travel-CoT) and an agentic architecture tailored for real-world travel assistance.

Travel-CoT: Structured Reasoning Enhancement

Pre-trained VLMs answer factual queries well but struggle with planning queries that require multi-factor reasoning. To address this, Travel-CoT explicitly decomposes each query into spatial, temporal, and practical reasoning steps. Following reviewer feedback, we adopt a two-stage formulation

Method	LLM	Image Size	Sample Size	Pure Text Score	Pure Text Δ	VQA Score	VQA Δ	Full Score	Full Δ
Pretrained Models									
BLIP-2 (Li et al. 2023a)	Vicuna-13B (Chiang et al. 2023)	224^2	129M	60.3	-	51.6	-	56.9	-
InstructBLIP (Dai et al. 2023)	Vicuna-7B (Chiang et al. 2023)	224^2	129M	62.8	-	54.1	-	59.4	-
InstructBLIP (Dai et al. 2023)	Vicuna-13B (Chiang et al. 2023)	224^2	129M	64.6	-	55.4	-	61.1	-
Shikra (Chen et al. 2023b)	Vicuna-13B (Chiang et al. 2023)	224^2	600K	71.6	-	60.8	-	67.5	-
Qwen-VL (Bai et al. 2023b)	Qwen-7B (Bai et al. 2023a)	448^2	1.4B	72.1	-	61.6	-	68.1	-
Qwen-VL-Chat (Bai et al. 2023b)	Qwen-7B (Bai et al. 2023a)	448^2	1.4B	73.2	-	62.8	-	69.2	-
LLaVA-1.5 (Liu et al. 2023a)	Vicuna-7B (Chiang et al. 2023)	336^2	558K	72.8	-	62.3	-	68.8	-
LLaVA-1.5 (Liu et al. 2023a)	Vicuna-13B (Chiang et al. 2023)	336^2	558K	74.3	-	63.3	-	70.0	-
Fine-tuned Models									
BLIP-2 (Li et al. 2023a)	Vicuna-13B (Chiang et al. 2023)	224^2	129M	64.7	**+7.3%**	54.7	**+6.0%**	60.9	**+7.0%**
InstructBLIP (Dai et al. 2023)	Vicuna-7B (Chiang et al. 2023)	224^2	129M	68.2	**+8.6%**	58.2	**+7.6%**	64.4	**+8.4%**
InstructBLIP (Dai et al. 2023)	Vicuna-13B (Chiang et al. 2023)	224^2	129M	68.8	**+6.5%**	58.8	**+6.1%**	64.9	**+6.2%**
Shikra (Chen et al. 2023b)	Vicuna-13B (Chiang et al. 2023)	224^2	600K	77.7	**+8.5%**	66.7	**+9.7%**	73.5	**+8.9%**
Qwen-VL (Bai et al. 2023b)	Qwen-7B (Bai et al. 2023a)	448^2	1.4B	78.7	**+9.2%**	67.7	**+9.9%**	74.5	**+9.4%**
Qwen-VL-Chat (Bai et al. 2023b)	Qwen-7B (Bai et al. 2023a)	448^2	1.4B	78.4	**+7.1%**	67.4	**+7.3%**	74.2	**+7.2%**
LLaVA-1.5 (Liu et al. 2023a)	Vicuna-7B (Chiang et al. 2023)	336^2	558K	78.0	**+7.1%**	67.0	**+7.5%**	73.8	**+7.3%**
LLaVA-1.5 (Liu et al. 2023a)	Vicuna-13B (Chiang et al. 2023)	336^2	558K	80.4	**+8.2%**	68.9	**+8.8%**	76.0	**+8.6%**
Fine-tuned Models with Travel-CoT									
TraveLLaMA (Ours)	Vicuna-13B (Chiang et al. 2023)	336^2	558K	82.5	**+10.7%**	70.5	**+11.0%**	77.8	**+10.8%**

Table 3: Comparisons of performance scores (Pure Text, VQA, and Full scores) and improvements across different vision-language models, showing both pretrained and fine-tuned versions with their respective LLM backbones, image sizes, and sample sizes. Full score is computed as weighted average based on test set distribution (61.5% text, 38.5% vision-language).

that better reflects our true implementation. Given multimodal input (x, Q), the model first generates a reasoning chain

$$r = f_\theta(x, Q), \quad r = \{r_s, r_t, r_p\},$$

where r_s encodes spatial understanding (locations, distances, routes), r_t encodes temporal scheduling (operating hours, time allocation), and r_p captures practical constraints (budget, accessibility, safety). The final answer is then generated conditioned on both the input and the reasoning chain:

$$y \sim P_\phi(y \mid x, Q, r).$$

We train the two components jointly using supervised learning on 5,000 expert-annotated Travel-CoT examples:

$$\mathcal{L} = \lambda\, \mathcal{L}_{\mathrm{CoT}}(r^*, r) + \mathcal{L}_{\mathrm{ans}}(y^*, y).$$

This structured reasoning improves plan correctness and interpretability while maintaining training efficiency through parameter-efficient tuning.

Agent Architecture for Real-time Planning

For practical deployment, we integrate Travel-CoT into a ReAct-style agent capable of processing multimodal travel requests with iterative reasoning and tool use (Figure 4). The agent proceeds in four stages. First, *Query Analysis* extracts textual constraints (destination, days, budget, group size) and interprets visual inputs such as uploaded photos for landmark recognition. Second, *Reasoning* applies Travel-CoT to organize spatial, temporal, and practical requirements into an actionable planning state.

Next, the agent enters *Tool Employment*, calling APIs for schedules, prices, reviews, and transit information. Its internal state evolves as:

$$\mathrm{Plan}_t = \mathrm{Update}\big(\mathrm{Plan}_{t-1},\ \mathrm{Tool}(\pi(s_t, \mathrm{Plan}_{t-1})),\ r\big),$$

where π selects appropriate tools based on the current reasoning state. Returned information (e.g., hours, prices, availability) is merged with visual grounding from the input images.

Finally, the agent executes *Result Integration*, generating the final plan:

$$y = \mathrm{Generate}(\mathrm{Plan}_T, \{o_1, \ldots, o_T\}, r),$$

producing detailed itineraries with schedules, budgets, and constraint checks. The agent verifies compliance with user requirements (e.g., total cost, hours, accessibility) before returning the answer. For example, given an uploaded image of the Brooklyn Bridge and a budget limit, the agent identifies the landmark, retrieves hours and nearby attractions, and synthesizes a full-day schedule with cost calculations and timing alignment.

Experiments

Experimental Setup

Dataset Configuration. We divide TravelQA into training and testing sets with a 4:1 ratio, ensuring all splits are POI-disjoint across text, images, and metadata to prevent leakage. For evaluation, each test question is converted into a four-choice multiple-choice (MCQ) format to obtain precise, reproducible accuracy metrics. Distractors are constructed from semantically plausible alternatives within the same category or region (e.g., nearby POIs, similar attractions), rather than random sampling. Models generate free-form answers, which are mapped to MCQ options using normalized string matching and a lightweight semantic matcher; responses failing to match any option are counted as incorrect. Our assessment spans pure text understanding (travel queries, cultural knowledge), map comprehension for spatial layout and navigation cues, scene understanding for visual

A. Overall SUS Scores			
	System	**SUS Score**	**Rating**
	Ours	**82.5**	**Excellent**
	Claude 3.5	76.3	Good
B. Items (0-10): Odd=(score-1)×2.5, Even=(5-score)×2.5			
Item	**Description**	**Ours**	**Claude 3.5**
1	Would use frequently ↑	**8.8**	8.0
2	Unnecessarily complex ↓	**8.5**	7.3
3	Easy to use ↑	**8.6**	7.8
4	Need technical support ↓	**8.8**	7.5
5	Functions well integrated ↑	**8.0**	7.7
9	Confident using ↑	7.9	**8.0**
6	Too much inconsistency ↓	**8.3**	6.9
10	Needed to learn a lot ↓	7.9	**7.8**
7	Quick to learn ↑	**8.7**	8.1
8	Cumbersome to use ↓	**8.0**	7.2
C. Supplementary (1-5)			
Item	**Question**	**Ours**	**Claude 3.5**
11	Recommendations met needs	**4.5**	4.1
12	Time Conservation	**4.6**	4.2
13	Trust in recommendations	4.3	**4.4**
14	Would recommend to others	**4.5**	4.0
D. Participant Demographics			
Total participants		500 (250 per system)	
Age range		18-62 years	

Table 4: System Usability Scale (SUS) Evaluation Results.

and architectural attributes, information extraction for operational details, review comprehension for synthesizing user-generated opinions, and temporal reasoning for operating hours and seasonal constraints. This comprehensive MCQ-based protocol provides a robust and standardized evaluation across all aspects of travel assistance.

Training Details. The training process follows a rigorous protocol designed to maximize model performance while ensuring reproducibility using 8 A100 GPUs. Our training dataset comprises 213k carefully curated QA pairs, representing a diverse range of travel-related scenarios and query types. The test set, consisting of 52k QA pairs, maintains a balanced distribution across different question categories and difficulty levels. We use normal QA for finetuning initally and then ues the CoT QA for post-training. All visual inputs are processed at a standardized resolution of 336 × 336 pixels, chosen to optimize the balance between computational efficiency and retention of important visual details. Text inputs are limited to a maximum length of 512 tokens, ensuring comprehensive coverage of complex travel queries while maintaining practical processing efficiency. The training process implements standard optimization techniques, including learning rate scheduling, gradient clipping, and early stopping based on validation performance.

Evaluation Tasks. Our evaluation combines quantitative QA accuracy, large-scale human usability assessment, and qualitative analysis. For quantitative testing, we evaluate TraveLLaMA on 52k POI-disjoint test questions in the TravelQA benchmark, covering both text-only queries (cultural information, travel advice) and vision–language tasks (landmark recognition, scene interpretation, and map-based spatial reasoning). Each test item is formatted as a four-choice MCQ, enabling precise measurement of model correctness. To assess real-world usability, we conduct a large-scale user study with 500 participants using the System Usability Scale (SUS). The study adopts a between-subjects design with anonymized system identities (blind evaluation). Participants complete four multi-step travel-planning tasks—including itinerary construction, constraint satisfaction, and iterative refinement—and then rate usability on the 10-item SUS questionnaire. We follow the standard SUS scoring procedure (1–5 Likert transformed to 0–100) and report results with 95% confidence intervals, effect size (Cohen's d), and statistical significance using a two-sided t-test. Finally, we perform qualitative comparisons between TraveLLaMA and strong baselines across representative travel scenarios, examining factual accuracy, constraint adherence, and practical utility.

Performance Comparison. Our experimental results in Table 3 demonstrate substantial improvements through domain-specific fine-tuning. Among pretrained models, LLaVA-1.5 with Vicuna-13B achieves the highest baseline performance (Pure Text: 74.3, VQA: 63.3, Full: 70.0). Fine-tuning on TravelQA yields consistent gains across all architectures, with improvements ranging from +6.2% to +9.4%. Notably, Qwen-VL shows the largest relative improvements (+9.2% Pure Text, +9.9% VQA, +9.4% Full), while LLaVA-1.5 (Vicuna-13B) achieves the best absolute performance among fine-tuned baselines (80.4, 68.9, 76.0). Our complete TraveLLaMA system, incorporating Travel-CoT reasoning, further enhances performance to 82.5 Pure Text (+10.7%), 70.5 VQA (+11.0%), and 77.8 Full (+10.8%) compared to the pretrained baseline. The results also reveal that models with larger image resolutions (448^2 for Qwen-VL vs 224^2 for BLIP-2) achieve better performance.

Agent Performance Analysis. Table 4 shows that TraveLLaMA achieves markedly higher usability than Claude3.5, with an SUS score of 82.5 ("Excellent") versus 76.3 ("Good"). The 6.2-point gain is driven by TraveLLaMA's domain-optimized design, reflected in strong ease-of-use, learnability, and reduced-complexity ratings. Users consistently found TraveLLaMA more intuitive and less cognitively demanding for travel planning. Claude3.5 shows small advantages in confidence and trust, likely influenced by general familiarity with the model. Based on a large-scale study of 500 users, the results indicate that a specialized travel agent with structured reasoning (Travel-CoT) provides a significantly better user experience than a general-purpose chatbot, particularly in minimizing cognitive load and improving task efficiency.

Qualitative Analysis. Figure 5 illustrates that TraveLLaMA consistently outperforms Claude3.5 in accuracy and contextual understanding across diverse travel-related queries. In map-based tasks, TraveLLaMA shows precise location grounding—for example, correctly identifying Piazza Sidney in Rome's Trastevere district and providing relevant nearby information such as local restaurants and operating hours—while Claude3.5 often fails to recognize the correct place. In scene recognition, TraveLLaMA accurately identifies iconic landmarks like Shanghai's Nanjing Road and offers meaningful contextual descriptions, whereas Claude3.5

Figure 5: Comparison between TraveLLaMA and Claude 3.5 shows that TraveLLaMA provides more accurate and detailed travel information in these location-based examples.

mislabels it as Wangfujing Street in Beijing. TraveLLaMA also demonstrates stronger multimodal integration, delivering detailed, actionable establishment information for venues such as the Hong Kong Park Sports Centre and Quanjude Restaurant, where Claude3.5 gives incomplete or incorrect details. The performance gap widens in more complex queries requiring both visual reasoning and domain knowledge: TraveLLaMA not only recognizes architectural styles and venue-specific constraints but also proactively provides helpful travel guidance, including nearby attractions or optimal visiting times. These observations highlight the advantages brought by domain-specific tuning and structured reasoning in TraveLLaMA.

Conclusion

We presented TraveLLaMA, a specialized multimodal language model that advances AI-powered travel assistance through three key contributions: (1) TravelQA, the first large-scale dataset with 265k QA pairs combining text, images, and maps for travel-specific training; (2) Travel-CoT, a structured reasoning framework that decomposes queries into spatial, temporal, and practical dimensions, achieving 10.8% accuracy improvement while providing interpretable decision paths; and (3) an interactive agent system that integrates real-time services for dynamic planning, validated through user studies with 500 participants achieving a SUS score of 82.5. Our comprehensive experiments demonstrate that domain-specific training yields 6.2-9.4% improvements over general-purpose models, with Travel-CoT providing substantial additional gains. Beyond establishing effective methods for travel AI, this work offers valuable insights for developing specialized assistants in other complex multimodal domains. Future research directions include real-time information integration, multilingual support, and personalized planning mechanisms to further enhance the travel experience.

Acknowledgements

This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

References

Andreas, J. 2022. Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Azerbayev, Z.; Schoelkopf, H.; Paster, K.; Santos, M. D.; McAleer, S.; Jiang, A. Q.; Deng, J.; Biderman, S.; and Welleck, S. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *CoRR*, abs/2308.12966.

Brown, S.; et al. 2023. Limitations in Human-Level Understanding of Visual Travel Content. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chen, H.; Pasunuru, R.; Weston, J.; and Celikyilmaz, A. 2023a. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *CoRR*, abs/2306.15195.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chu, M.; Li, Y.; and Chua, T.-S. 2025. GraphVideoAgent: Enhancing Long-form Video Understanding with Entity Relation Graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 4639–4648.

Chu, M.; Zheng, Z.; Ji, W.; Wang, T.; and Chua, T.-S. 2024. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, 213–231. Springer.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR*.

Deng, C.; Zhang, T.; He, Z.; Chen, Q.; Shi, Y.; Xu, Y.; Fu, L.; Zhang, W.; Wang, X.; Zhou, C.; et al. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 161–170.

Feng, J.; Du, Y.; Liu, T.; Guo, S.; Lin, Y.; and Li, Y. 2024. Citygpt: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*.

Frome, A.; et al. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *Advances in Neural Information Processing Systems*.

Ge, Y.; Hua, W.; Mei, K.; Tan, J.; Xu, S.; Li, Z.; Zhang, Y.; et al. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36: 5539–5568.

Godey, N.; de la Clergerie, É.; and Sagot, B. 2024. On the scaling laws of geographical representation in language models. *arXiv preprint arXiv:2402.19406*.

Gurnee, W.; and Tegmark, M. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Kiros, R.; et al. 2014. Unifying Visual-Semantic Embeddings. *arXiv preprint arXiv:1411.2539*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Li, L.; Ye, Y.; Jiang, B.; and Zeng, W. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.

Li, Z.; Zhou, W.; Chiang, Y.-Y.; and Chen, M. 2023b. Geolm: Empowering language models for geospatially grounded language understanding. *arXiv preprint arXiv:2310.14478*.

Liang, X.; Wang, B.; Huang, H.; Wu, S.; Wu, P.; Lu, L.; Ma, Z.; and Li, Z. 2023. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. *CoRR*, abs/2304.08485.

Liu, Y.; et al. 2023b. LLaVA: Large Language and Vision Aligned Model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lu, P.; Peng, B.; Cheng, H.; Galley, M.; Chang, K.-W.; Wu, Y. N.; Zhu, S.-C.; and Gao, J. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36: 43447–43478.

Maheshwary, S.; Paul, A.; and Sohoney, S. 2024. Pretraining and finetuning language models on geospatial networks for accurate address matching. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 763–773.

Manvi, R.; Khanna, S.; Mai, G.; Burke, M.; Lobell, D.; and Ermon, S. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Momennejad, I.; Hasanbeig, H.; Vieira Frujeri, F.; Sharma, H.; Jojic, N.; Palangi, H.; Ness, R.; and Larson, J. 2023. Evaluating cognitive maps and planning in large language models with cogeval. *Advances in Neural Information Processing Systems*, 36: 69736–69751.

Nakajima, Y. 2023. Task-driven autonomous agent utilizing gpt-4, pinecone, and langchain for diverse applications. *See https://yoheinakajima. com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications (accessed 18 April 2023)*, 2: 3.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2023. GPT-4V: Visual Capabilities in GPT-4.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Schumann, R.; Zhu, W.; Feng, W.; Fu, T.-J.; Riezler, S.; and Wang, W. Y. 2024. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18924–18933.

Shafik, W. 2024. Introduction to ChatGPT. In *Advanced applications of generative AI and natural language processing models*, 1–25. IGI Global.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36: 38154–38180.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36: 8690–8701.

Wang, X.; Ling, X.; Zhang, T.; Li, X.; Wang, S.; Li, Z.; Zhang, L.; and Gong, P. 2023. Optimizing and fine-tuning large language model for urban renewal. *arXiv preprint arXiv:2311.15490*.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xie, J.; Liang, Y.; Liu, J.; Xiao, Y.; Wu, B.; and Ni, S. 2023a. Quert: Continual pre-training of language model for query understanding in travel domain search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5282–5291.

Xie, T.; Zhou, F.; Cheng, Z.; Shi, P.; Weng, L.; Liu, Y.; Hua, T. J.; Zhao, J.; Liu, Q.; Liu, C.; et al. 2023b. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*.

Xie, W.; et al. 2024. TravelPlanner: AI-assisted Multi-day Itinerary Planning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yamada, Y.; Bao, Y.; Lampinen, A. K.; Kasai, J.; and Yildirim, I. 2023. Evaluating spatial understanding of large language models. *arXiv preprint arXiv:2310.14540*.

Yang, H.; Yue, S.; and He, Y. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Yang, J.; Ding, R.; Brown, E.; Qi, X.; and Xie, S. 2024. V-irl: Grounding virtual intelligence in real life. In *European Conference on Computer Vision*, 36–55. Springer.

Yang, K.; et al. 2023. Specialized Architectures for Vision-Language Models in Tourism. *arXiv preprint arXiv:2306.10456*.

Yu, J.; Wu, Y.; Chu, M.; Ren, Z.; Huang, Z.; Chu, P.; Zhang, R.; He, Y.; Li, Q.; Li, S.; Li, Z.; Tu, Z.; He, C.; Qiao, Y.; Wang, Y.; Wang, Y.; and Wang, L. 2025. VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21655–21666.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhao, L.; et al. 2023a. Domain-Specific Pre-training for Vision-Language Models. *Proceedings of the International Joint Conference on Artificial Intelligence*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.

Zhou, W.; Jiang, Y. E.; Cui, P.; Wang, T.; Xiao, Z.; Hou, Y.; Cotterell, R.; and Sachan, M. 2023. Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*.