Supporting information for "SynopticBench: Evaluating Vision-Language Models on Generating Weather Forecast Discussions of the Future"

June 2026

Contents of this File

Implementation Details. We fine-tune each base model by training for roughly 1-3 epochs on one NVIDIA H200 GPU before running it on the test set. We use 2016-2022 for the training set, 2023 as the validation set, and 2024-2025 as the test set. The projector weights were fully fine-tuned, while Low-Rank Adaptation (LoRA) was used to train a small portion of the attention mechanism in the language model (<2%). We apply weight decay and dropout to the LoRA adapter branches to prevent overfitting. In addition, early stopping is used by computing Space-local scores for the validation set every 5000 steps.

Prompt Details. Base model and finetuned model versions for LLaVA-v1.5-7B, LLaVA-v1.5-13B, Qwen2-VL-7B, and LLaMA-3.2-11B-Vision were run on the test set. Each model uses the following prompt to generate discussions (see Fig. 1): "Analyze these weather charts showing maximum precipitation over the 3-day period (shaded contours), 500 mb geopotential height (unshaded contours) and 850 mb wind velocity (wind barbs) for the <[City, State]> forecast region and generate a detailed forecast discussion of the large-scale features relevant to the area. The forecast region is shown in the yellow box".

Additional Preprocessing Details. The images used in this study use the temporal mean of lead times ranging from 3 hours to 48 hours in 3 hour increments. The anomalies are calculated from monthly means ranging across the entire dataset. The month chose for climatology is always the same month that the AFD is issued. To create forecast images, 2-m temperature anomaly is capped between -5° and 5°C. 500 mb geopotential height contours are drawn at intervals of 60m and 850mb wind velocity barbs are plotted every 20th grid point to avoid visual clutter.

To filter AFDs, we only extract sentences containing the following keywords: *trough, trof, the low, this low, upper level low, low pressure, low-pressure, upper low, cyclone, closed low, cut-off low, troughing, ridge, the high, this high, upper level high, high pressure, high-pressure, upper high, anticyclone, blocking, ridging, cold front, warm front, warmer, cooler, and freezing*. We also exclude sentences with the following model keywords: *ECMWF, EURO, HRRR, ECCC, CMC, GEM, NAM, UKMET, ICON, RAP, SREF, HREF,* and synoptic pressure system keywords: *shortwave, short-wave, sfc trough, surface trough, sfc ridge, surface ridge, surface low, sfc low, surface high, sfc high* to increase the relevancy of NWS text in the use case presented in this study.

Parameter Settings. The same settings for the finetuning parameters were used for LLaVA-v1.5-7B, LLaVA-v1.5-13B, Qwen2-VL-7B, and LLaMA-3.2-11B-Vision. The learning rate was set at 2e-5, the batch size was set to 16, the LoRA rank was set to 8, and the LoRA alpha was set to 16. LoRA was applied to all linear modules within the attention and multi-layer perceptron blocks. AdamW was used as an optimizer. Image preprocessing was handled natively by each model's respective Hugging Face AutoProcessor. The images were converted to RGB and dynamically resized, padded, and normalized according to the base model's default vision encoder requirements (e.g., SigLIP for LLaVA/Llama and dynamic resolution for Qwen2.5-VL) before being tokenized.

Data Quality Control. Each text sample is matched to the closest GFS run. The initial raw dataset contained 1,367,041 samples. Text samples that are not within 12 hours of any GFS run are discarded. Pre-filtered samples in the raw dataset containing less than 30 words or greater than 200 words were also discarded. In addition, we removed samples containing text that suggested longer lead times beyond the 48-hour lead time post filtering (e.g. "Day 3", "extended period", "long term"). Quality control tests revealed zero samples containing text duplicates, unusual characters, absent image tokens, missing image files, or corrupted image files. Once all of the filtering specific to the example use case of this experiment was completed, the training set was reduced to 457,230 high-quality samples.

Extraction Details. We developed an automated Python pipeline to interface with

the IEM JSON API, systematically querying and downloading historical AFDs from 117 NWS Weather Forecast Offices. The raw text files were extracted for the period spanning January 2019 to November 2025, yielding a comprehensive collection of chronologically ordered NWS discussions. To ensure strict temporal accuracy for multimodal pairing, we implemented a custom regular expression parser to extract the exact issuance timestamp directly from the raw text body of each product, converting all times to Coordinated Universal Time (UTC).

LLM as a Judge. Gemini-2.5-Flash was used as a judge for evaluation. The following text was used as the prompt for the judge: "Compare the generated forecast discussion to the ground truth forecast discussion. Evaluate how accurately the generated discussion captured the synoptic-scale features. Ignore minor differences in phrasing and focus on meteorological accuracy and completeness. Score the prediction from 0 to 1:

0: Completely incorrect or severely hallucinated.

.25: Major meteorological errors or missing critical synoptic features.

.5: Partially correct, but missing some context or containing minor errors.

.75: Mostly accurate and aligns well with the ground truth, minor omissions.

1: Excellent, highly accurate, and meteorologically sound compared to the ground truth.

Location Stop Nodes. The following locations are excluded from functioning as common relatives or matching to other locations via a common relative: "Canada", "CONUS", "Eastern Canada", "Central Canada", "Western Canada", "Eastern CONUS", "Western CONUS", "Central CONUS", "Central Plains", "Ohio Valley", "Great Lakes", "Central U.S.", "Eastern U.S.", "Western U.S.", "Central United States", "Eastern United States", "Western United States", "Central U", "Eastern U", "Western U", "Eastern US", "Western US", "Central US", "Midwest", "The Plains".

SPACE Temperature Scores. SPACE scores for temperature are shown in Table 4. The following key terms for temperature polarity were used: "warm front", "warmer than average", "warmer than normal", "above average temperature", "cold front", "cooler than average", "cooler than normal", "colder than normal", and "below average temperature". The number of temperature terms found in both AFDs and generated text was dominated by "cold front" and "warm front" (>80%). Fine-tuned Qwen2.5-VL-7B-LoRA had the highest SPACE-local scores while fine-tuned LLaVA-v1.5-7B had the highest SPACE-aggregate scores. The 48-hour temporal average that was used to create the images could lead to too much smoothing for the temporal scale at which fronts often take effect. Although SPACE scores could be effective for synoptic temperature systems, altering training images and experimental setup to target temperature-related phenomena would improve the reliability of evaluation in future work.

Additional Example Cases. Three additional example cases are shown in Figures A1-A3. In Case 3, there are two matching instances of pressure systems and two unmatching instances of pressure systems, leading to a coverage ratio of 0.5. Of the matching instances, both indicated high pressure, resulting in a match score of 1 and a SPACE-local score of 0.5. In Case 4, the prediction described synoptic pressure systems, while the reference text did not (Short wave troughs are not considered synoptic pressure systems). This results in a SPACE-local score that defaults to 0. Although there was some accurate information in the generated text, it was not relevant enough to the area to be included in the AFD. The pressure systems described in the generated

Model	s_s^{loc}	s_m^{loc}	r_c^{loc}	s_s^{agg}	s_m^{agg}	r_c^{agg}
LLaVA-v1.5-7B-base	.0173$_{\pm.0028}$	.8278$_{\pm.0526}$	.0207$_{\pm.0031}$	.3447$_{\pm.0028}$	.5602$_{\pm.0526}$	.6017$_{\pm.0031}$
LLaVA-v1.5-13B-base	.0793$_{\pm.0048}$	**.9068**$_{\pm.0125}$	.0862$_{\pm.0051}$	.4229$_{\pm.0008}$	.6053$_{\pm.0056}$	.6843$_{\pm.0013}$
LLaMA-3.2-11B-Vision-base	.0671$_{\pm.0025}$	.7294$_{\pm.0075}$	.0886$_{\pm.0030}$	.3300$_{\pm.0025}$	.5683$_{\pm.0075}$	.5709$_{\pm.0030}$
Qwen2.5-VL-7B-base	.0491$_{\pm.0032}$	.8330$_{\pm.0155}$	.0577$_{\pm.0035}$	.3302$_{\pm.0032}$	.5478$_{\pm.0155}$	.5878$_{\pm.0035}$
LLaVA-v1.5-7B-LoRA	.1863$_{\pm.0056}$	.8839$_{\pm.0075}$	.2077$_{\pm.0060}$	**.4679**$_{\pm.0056}$	.5965$_{\pm.0075}$	.7718$_{\pm.0060}$
LLaVA-v1.5-13B-LoRA	.1862$_{\pm.0063}$	.8697$_{\pm.0085}$	.2098$_{\pm.0067}$	.4209$_{\pm.0063}$	.5870$_{\pm.0085}$	.7043$_{\pm.0067}$
LLaMA-3.2-11B-Vision-LoRA	.1726$_{\pm.0055}$	.8992$_{\pm.0073}$	.1892$_{\pm.0057}$	.4597$_{\pm.0055}$	**.6124**$_{\pm.0073}$	.7395$_{\pm.0057}$
Qwen2.5-VL-7B-LoRA	**.1886**$_{\pm.0074}$	.8660$_{\pm.0058}$	**.2128**$_{\pm.0062}$	.3992$_{\pm.0074}$	.5978$_{\pm.0058}$	.6529$_{\pm.0062}$
Gemini-3.1-Pro	.0397$_{\pm.0050}$	.7725$_{\pm.0340}$	.0408$_{\pm.0057}$	.3963$_{\pm.0050}$	.5135$_{\pm.0340}$	**.7732**$_{\pm.0057}$
Nearest Neighbor	.0585$_{\pm.0036}$	.7532$_{\pm.0166}$	.0753$_{\pm.0041}$	.3216$_{\pm.0036}$	.5607$_{\pm.0166}$	.5614$_{\pm.0041}$
Climatology	.0850$_{\pm.0042}$	.7728$_{\pm.0148}$	.1073$_{\pm.0048}$	.4042$_{\pm.0042}$	.5795$_{\pm.0148}$	.6894$_{\pm.0048}$
Llama-3.2 (blind)	.0912$_{\pm.0048}$	.8664$_{\pm.0141}$	.1043$_{\pm.0051}$	.4211$_{\pm.0048}$	.5885$_{\pm.0141}$	.7050$_{\pm.0051}$

Table 1: SPACE was used for synoptic temperature system evaluation for base models, finetuned models, and baselines. We calculate the mean and standard error of the SPACE-local and SPACE-aggregate scores (s_s), match scores (s_m), and coverage ratios (r_c) within the test set. The range of possible scores for all metrics is 0 to 1, with 1 being a perfect score

text could still impact SPACE-aggregate scores. In Case 5, all pressure systems were described in the same general area, leading to a coverage ratio of 1. The match score was imperfect due to a ridge that would impact the region soon after the trough.

Training Set Size. The impact of the size of the training set on the SPACE scores is shown in Figure A4. The majority of the improvement in SPACE scores occurs within the first half epoch of training.

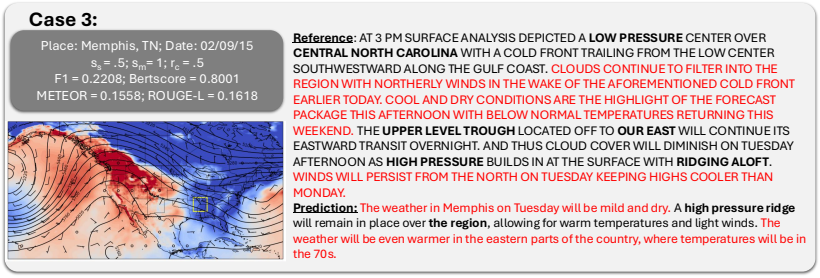
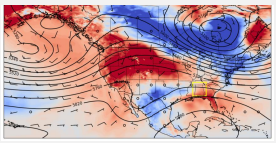


Figure A1: Example Case 3

Case 4:

Place: Birmingham, AL; Date: 03/20/15
$s_s = 0$; $s_m = 0$; $r_c = 0$
F1 = 0.0667; Bertscore = 0.8250
METEOR = 0.0779; ROUGE-L = 0.0811



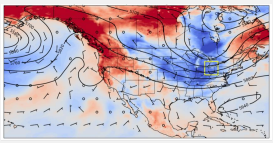
Reference: WE HAVE A SOUTHERN STREAM SHORT WAVE TROUGH MOVING ACROSS THE AREA.

Prediction: As the upper level **trough** moves to **the east**, the low level flow will shift to the east-southeast. The surface flow will be southerly over the next couple of days, with a **low pressure** area developing over the **eastern Gulf of Mexico**. **High pressure** will build over the **western United States**. A **ridge** of **high pressure** will build over **the Southeast**.

Figure A2: Example Case 4

Case 5:

Place: Cleveland, OH; Date: 08/03/17
$s_s = .67$; $s_m = .67$; $r_c = 1$
F1 = 0.1942; Bertscore = 0.8250
METEOR = 0.1253; ROUGE-L = 0.2149



Reference: **Low pressure** from the upper **midwest** will move across the **Great lakes** Friday pulling a strong cold front east across the area Friday. **High pressure** will build across **the region** over the weekend.

Prediction: The latest forecast discussion for the large-scale features indicates that a **low pressure** system will continue to develop and track across the eastern **Great Lakes** this afternoon and evening. **This low** will deepen and intensify as it moves over the colder waters of **Lake Superior** tonight. **The low** will bring gusty winds and rain to **the region**, with a potential for thunderstorms and heavy rain. The winds will be strongest in the morning and afternoon hours, with gusts over 40 mph expected across the eastern Great Lakes.

Figure A3: Example Case 5

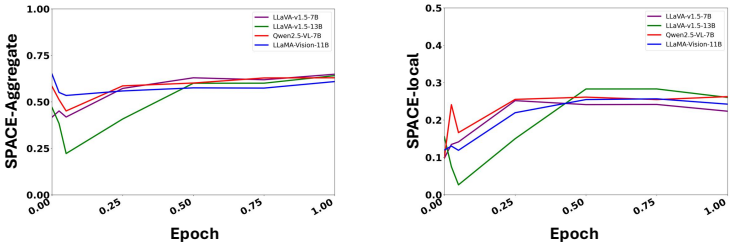


Figure A4: Training Set Size