# Test-Time Consistency in Vision Language Models Supplementary Material

Shih-Han Chou*[1,2], Shivam Chandhok*[1,2], James J. Little[1], Leonid Sigal[1,2,3]
[1]University of British Columbia, Canada
[2]Vector Institute for AI, Canada [3]Canada CIFAR AI Chair, Canada
{shchou75, chshivam, little, lsigal}@cs.ubc.ca

## 1. Ablation on Decoding Temperature

Table 1. **Different temperature on LLaVa-Next.**

| Temp | Acc | $S_{GT}$ | Con | $S_C$ | $O_{all}$ |
|---|---|---|---|---|---|
| **Question Rephrasing** | | | | | |
| LLaVa-NEXT, $\tau = 0$ | 42.89 | 64.89 | 49.18 | 65.69 | 55.61 |
| + Constant $T$ | 44.48 | 68.74 | 83.39 | 88.47 | **68.25** |
| LLaVa-NEXT, $\tau = 0.5$ | 42.06 | 65.29 | 52.02 | 66.52 | 56.33 |
| + Constant $T$ | 44.48 | 68.74 | 83.39 | 88.47 | **68.25** |
| LLaVa-NEXT, $\tau = 1$ | 42.06 | 65.29 | 52.02 | 66.52 | 56.33 |
| + Constant $T$ | 44.48 | 68.74 | 83.39 | 88.47 | **68.25** |
| **Image Restyling** | | | | | |
| LLaVa-NEXT, $\tau = 0$ | 17.57 | 41.47 | 55.34 | 71.36 | 40.27 |
| + Constant $T$ | 18.99 | 42.49 | 88.25 | 91.25 | **45.80** |
| LLaVa-NEXT, $\tau = 0.5$ | 17.57 | 41.47 | 55.34 | 71.36 | 40.27 |
| + Constant $T$ | 17.64 | 40.64 | 82.80 | 76.64 | **42.68** |
| LLaVa-NEXT, $\tau = 1$ | 17.57 | 41.47 | 55.34 | 71.36 | 40.27 |
| + Constant $T$ | 17.64 | 40.64 | 82.80 | 76.64 | **42.68** |
| **Context Reasoning** | | | | | |
| LLaVa-NEXT, $\tau = 0$ | 30.24 | 27.43 | 32.11 | 58.44 | 35.23 |
| + Constant $T$ | 32.50 | 50.84 | 89.91 | 90.16 | **56.97** |
| LLaVa-NEXT, $\tau = 0.5$ | 30.07 | 51.99 | 52.09 | 66.68 | 48.53 |
| + Constant $T$ | 32.31 | 53.84 | 93.4 | 95.31 | **59.15** |
| LLaVa-NEXT, $\tau = 1$ | 30.07 | 51.99 | 52.09 | 66.68 | 48.53 |
| + Constant $T$ | 32.31 | 53.84 | 93.4 | 95.31 | **59.15** |

We conduct an ablation to assess the impact of decoding temperature $\tau$ on our test-time consistency framework using LLaVA-NEXT across three perturbation types: *Question Rephrasing*, *Image Restyling*, and *Context Reasoning*.

Across all perturbations, our method improves consistency and overall robustness regardless of the temperature setting (see Table 1). Notably:

- **Question Rephrasing:** Our test-time strategy consistently boosts performance to a peak $O_{all} = 68.25$ for all values of $\tau$, indicating stable performance across decoding scales and strong resilience to linguistic variations.
- **Image Restyling:** While baseline performance is lower due to visual perturbations, our method still yields significant improvements. The best result is at $\tau = 0$, where $O_{all}$ improves from 40.27 to 45.80, a gain of 5.5 points.
- **Context Reasoning:** This task benefits most from our consistency framework. The best performance, $O_{all} = 59.15$, is achieved at both $\tau = 0.5$ and $\tau = 1$, indicating that our method improves reasoning-heavy tasks.

These results demonstrate that our approach is robust to temperature variation.

## 2. Efficiency & Practicality of Test-Time Adaptation

Our method introduces a small inference-time overhead, limited to 2 gradient updates on a single test sample and restricted to a lightweight subset of parameters (only the LM head). Table 2 shows a comparison with previous work in terms of training/inference time and parameters tuned. We believe this is a *worthwhile trade-off* given the overall efficiency it offers: it requires no large-scale task-specific training on curated data, adapts to new data distributions using just a single sample, and remains fully model-agnostic. On LLaVA, the average inference time increases modestly from 9.2 to 30.1 seconds per sample, an acceptable cost given the substantial gains in consistency and much less training time (18 hours vs 0) and parameters tuned (376M vs 131M). Gradient updates remain the dominant cost. Generating variants indeed incurs some cost (**0.22 sec** for K=4 style variants and **0.51 sec** for 3 rephrasings), but this process can be parallelized and only requires a single forward pass through a model (LLM or CNN), hence takes a fraction of time comparatively.

Table 2. **Efficiency Comparison between finetune and ours.** Training time in Our (test-time adaptation) is 0 denotes our method does not need training.

| Model | Training Time | Params Tuned | Inference Time |
|---|---|---|---|
| Finetune (MM-R$^3$[2]) | 18 hrs | 376 M | 9.2 s |
| Our (test-time adaptation) | 0 | 131 M | 30.1 s |

*Equal Contribution.

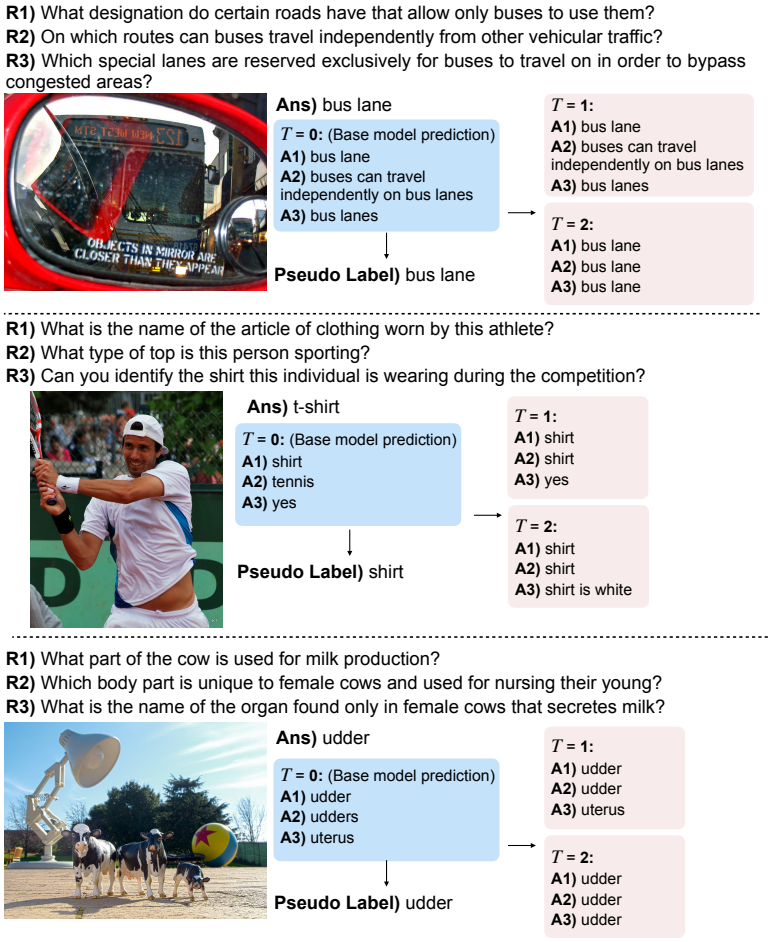


Figure 1. Qualitative results on the question rephrasing task.

## 3. Implementation Details.

We use the pre-trained VLMs as the base models and only fine-tune the language modelling head (LM-head) layer. We set updated steps $T = 2$ for the test-time experiments and maximum updated steps to $T = 4$ for the adaptive test-time experiments. The learning rate is set to $5e^{-4}$. All experiments are conducted on NVIDIA A40 with batch size 1 on all three models.

## 4. Evaluation Metrics

To systematically assess the performance of VLMs, we use four distinct evaluation metrics, on similar lines as previous work [2], each capturing different aspects of model performance.

**Accuracy (Acc).** To evaluate accuracy we assess the responses from VLMs based on an fuzzy string matching with the ground truth annotations, accounting for minor lexical variations. A similarity threshold of 85 is used to determine a match. The accuracy score is then calculated as the average of correct responses across the benchmark test-set.

**Similarity with GT ($\mathbf{S_{GT}}$).** Given the limitations of exact match criteria—which may penalize semantically correct responses for minor lexical differences—we employ a semantic similarity metric to better evaluate alignment between model outputs and ground truth. For example, terms like *person*, *man*, and *woman* are semantically related but would be treated as mismatches under strict accuracy metrics. To address this, we use BERT-based Sentence Similarity [3], which leverages contextual language model encodings to assess the semantic alignment between predictions and reference answers. This metric rewards semantic correctness over surface-form similarity. Final scores are computed as the average similarity across the dataset.

**Consistency Accuracy (Con).** This metric quantifies the proportion of responses that exhibit a predefined level of semantic consistency. We compute pairwise similarity scores between outputs using the same semantic similarity metric as in $S_{GT}$, and consider a pair consistent if its similarity exceeds a threshold of 0.7—motivated by observations from

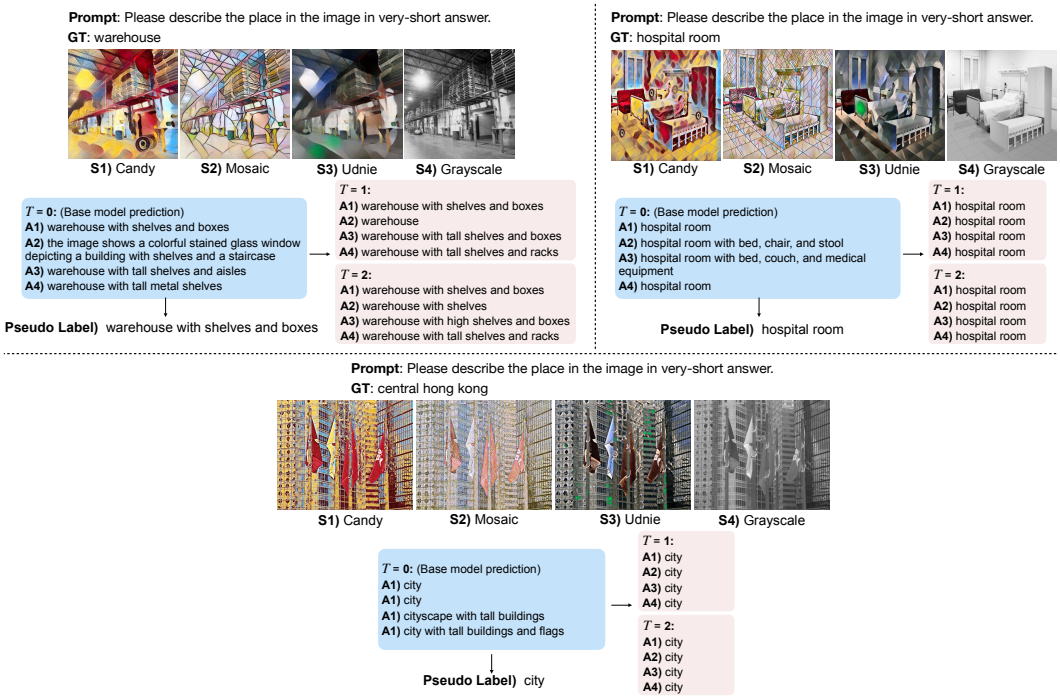


Figure 2. Qualitative results on the image restyling task.

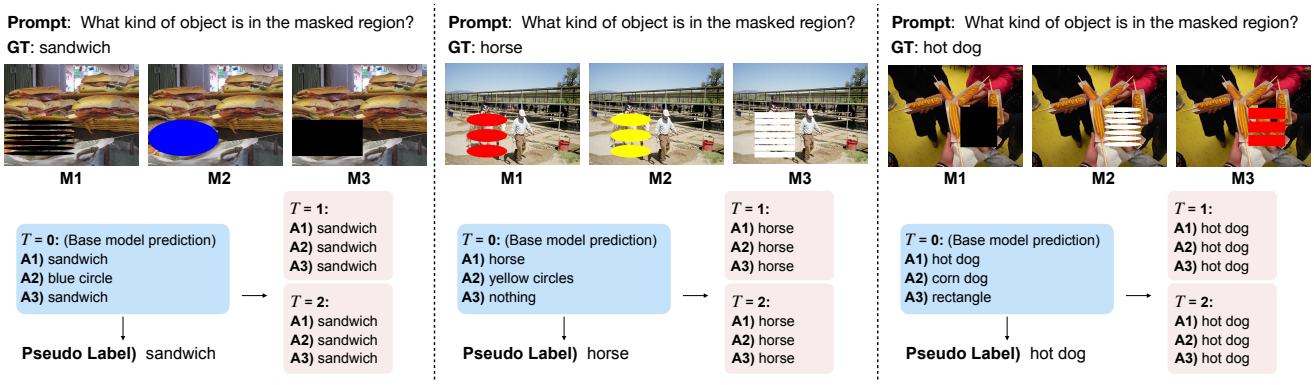


Figure 3. Qualitative results on the context reasoning task.

the Semantic Textual Similarity benchmark [1]. A response is deemed consistent if it meets this threshold with its paired counterpart.

The final score is calculated as the average proportion of consistent pairs across the dataset, providing an aggregate measure of the model's semantic stability across perturbed inputs.

**Consistency Similarity ($\mathbf{S_C}$).** Similar to the Consistency Accuracy metric, this measure computes pairwise semantic similarity scores between responses to assess consistency. However, instead of applying a threshold, we take the average of these similarity scores across the dataset. This provides a more *continuous* assessment of the model's coherence, capturing fine-grained variations in semantic consistency across perturbed inputs.

**Overall Performance ($O_{all}$).** We report overall model performance using the harmonic mean ($H_{\text{mean}}$) of correctness and consistency scores. Specifically, we first compute the average of $\mathbf{Acc}$ and $\mathbf{S_{GT}}$ to assess correctness, and the average of $\mathbf{Con}$ and $\mathbf{S_C}$ to assess consistency. These two averages are then combined using the harmonic mean:

$$H_{mean}(mean(\mathbf{Acc}, \mathbf{S_{GT}}), mean(\mathbf{Con}, \mathbf{S_C})). \quad (1)$$

We use the harmonic mean to balance correctness and consistency, as it penalizes models that perform well on only

**Questions for adaptation**

- What do you call a desk where people sit to apply makeup?
- What is the name of the furniture specifically designed for applying makeup?
- If a desk is used for sitting and putting on makeup, what is its common name?

**Questions for evaluation**

- What is the name of the furniture piece where individuals typically sit to apply their makeup?
- What is the term used to describe the desk specifically designed for makeup application?
- Can you provide three alternative ways to refer to the desk commonly used for putting on makeup?

---

**Questions for adaptation**

- What year did this team last secure a national championship?
- In which year did this team achieve its most recent national title victory?
- Can you tell me the year of this team's last national championship win?

**Questions for evaluation**

- In which year did this team most recently claim a national championship?
- What was the most recent year that this team tasted national title success?
- When was the last time this team emerged as national champions?

---

**Questions for adaptation**

- What do you call the part of the device that displays what you see and do?
- On this device, what is the name of the area that allows you to control what you see and do?
- The area on this device where you control your actions and what is displayed is known by what name?

**Questions for evaluation**

- What part of this device allows you to interact with its functions and display information?
- Can you identify the specific section of this device that enables you to manipulate its content and view?
- What is the name of the space on this device that allows you to control what is displayed and perform actions?

Figure 4. Examples of questions for adaptation and evaluation.

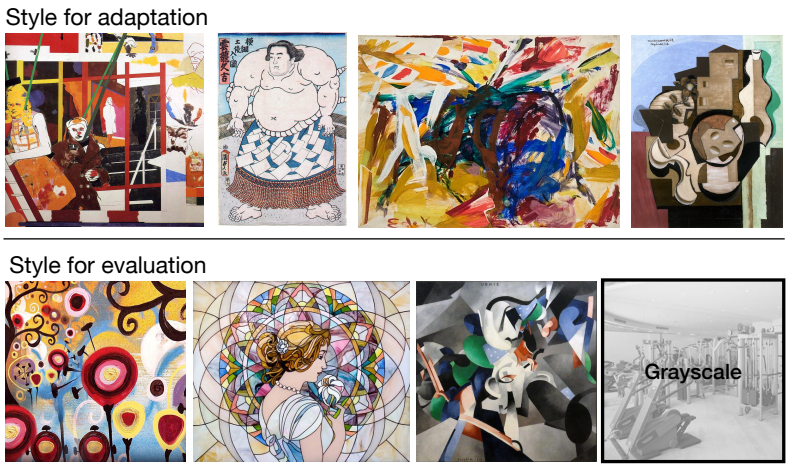


Figure 5. Examples of styled images for adaptation and evaluation.

one aspect, thereby encouraging robust performance across both dimensions.

## 5. Qualitative Results

We show qualitative results for the question rephrasing task in Figure 1, image restyling in Figure 2, and context reasoning in Figure 3. Across all three tasks, even when the base model predictions are inconsistent, our method is able to further improve consistency and thus overall score (as also supported by quantitative results in Main manuscript).

## 6. Examples for Disjoint Variant Setup.

Here we show samples used for our Disjoint Variant Setup which does adaptation and evaluation on separate sets of

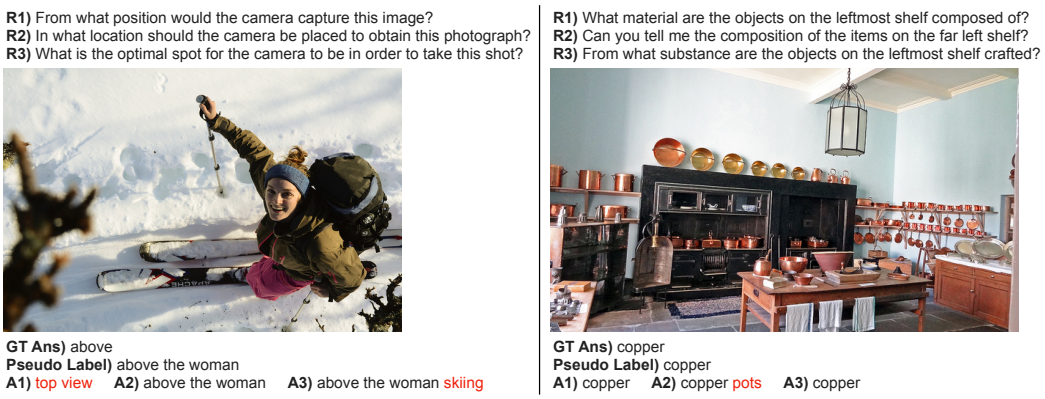


Figure 6. Failure results on the question rephrasing task.

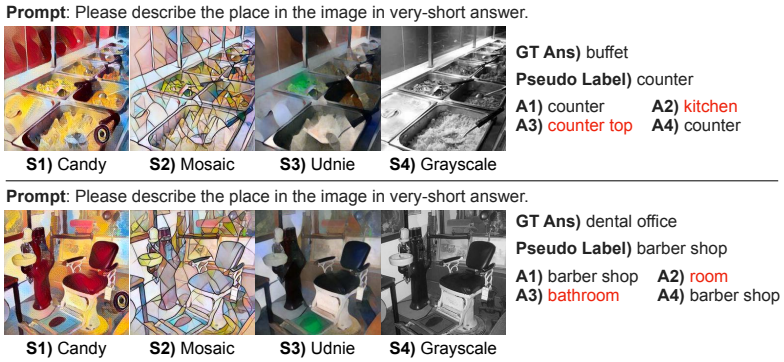


Figure 7. Failure results on the image restyling task.

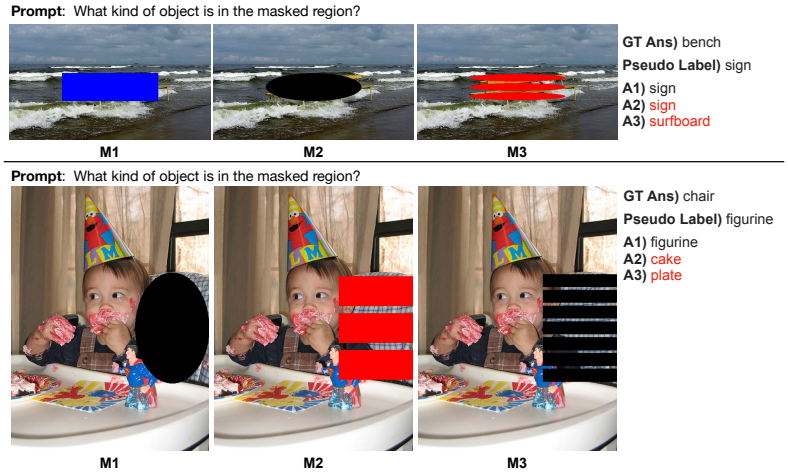


Figure 8. Failure results on the context reasoning task.

variants. For rephrasing we use a different LLM (Gemini vs. GPT4) and for restyling we restyle images using entirely different set of raw styles. We show examples of rephrased questions for adaptation and evaluation in Figure 4. In Figure 5, we provide the style images we used for the image restyling task for adaptation and evaluation.

## 7. Failure Cases

We show failure results for the question rephrasing task in Figure 6, image restyling in Figure 7, and context reasoning

in Figure 8.

## 8. Results on OKVQA

To further demonstrate generalization beyond MM-R$^3$, we additionally evaluated on OK-VQA dataset (Table 3). Results show that our method not only preserves, but often improves accuracy, while significantly improving consistency (*e.g.*, InternVL3: 55.33 $\rightarrow$ **91.73**; *Image Restyling*).

Table 3. **Results on OKVQA.**

| | Models | Acc | $S_{GT}$ | Con | $S_C$ | $O_{all}$ |
|---|---|---|---|---|---|---|
| Question Rephrasing | LLaVa-Next | 66.20 | 78.39 | 85.80 | 90.36 | 78.68 |
| | + Constant $T$ | 70.93 | 82.08 | 91.67 | 94.32 | 82.89 |
| | + Adapt. $T$ | 71.07 | 82.36 | 93.13 | 95.66 | **83.45** |
| | Qwen2-VL | 71.73 | 82.14 | 94.80 | 96.66 | 84.24 |
| | + Constant $T$ | 72.93 | 83.59 | 94.93 | 97.00 | 84.90 |
| | + Adapt. $T$ | 72.93 | 83.76 | 96.53 | 97.89 | **85.32** |
| | InternVL3 | 55.13 | 72.13 | 81.00 | 87.13 | 72.11 |
| | + Constant $T$ | 55.40 | 71.97 | 85.00 | 89.92 | 73.23 |
| | + Adapt. $T$ | 55.53 | 72.07 | 85.00 | 89.99 | **73.39** |
| Image Restyling | LLaVa-Next | 62.61 | 79.33 | 81.33 | 87.69 | 77.15 |
| | + Constant $T$ | 62.50 | 79.56 | 98.35 | 98.82 | 82.57 |
| | + Adapt. $T$ | 63.20 | 80.04 | 98.93 | 99.24 | **83.14** |
| | Qwen2-VL | 66.45 | 81.12 | 79.40 | 86.79 | 78.16 |
| | + Constant $T$ | 66.95 | 81.29 | 98.90 | 99.32 | 84.81 |
| | + Adapt. $T$ | 67.50 | 82.05 | 99.50 | 99.66 | **85.41** |
| | InternVL3 | 39.60 | 61.24 | 55.33 | 70.96 | 56.07 |
| | + Constant $T$ | 41.75 | 62.78 | 91.70 | 94.29 | 66.92 |
| | + Adapt. $T$ | 42.35 | 62.94 | 91.73 | 94.34 | **67.24** |

## References


[1] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics. 3

[2] Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. Mm-r$^3$: On (in-) consistency of multi-modal large language models (mllms). *ArXiv*, 2024. 1, 2

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 2