Supplementary Material
Learning by Hallucinating: Vision-Language Pre-training with Weak Supervision

Tzu-Jui Julius Wang, Jorma Laaksonen, Tomas Langer, Heikki Arponen, and Tom E. Bishop

1. Visualizing WFH-generated Representations

Figure 1 visualizes the hallucinated features along with other textual and visual representations. The hallucinated features appear to serve as the bridging representations across the V-L domains. Furthermore, the generated representations are contextual; for instance, *stage* and *actor* are close by and so are *artist*, *portrait*, *vector*, *blue*, and *beautiful*. This indicates that the proposed WFH capitalizes on the textual contextuality from VG object and attribute classes to generate reasonable contextual representations of the texts not present in VG, such as adjectives like *beautiful* and nouns like *vector*, *stage*, and *artist*.

2. Learning Visual Dictionary D

It is both crucial to learn D with quality representations and learn it efficiently. While there can be many ways to learn a good D, we opt for the simple K-means method with momentum updates, which is also used in [1]. Formally, given the randomly initialized visual words $\{\mathbf{d}_c \in \mathbb{R}^{2048}\}_{c=1}^{C}$, we update D with batches of $B_v = 512$ visual features $\{\mathbf{v}_i\}_{i=1}^{B_v}$ (pre-extracted by the same object detector that generates object and attribute tags on 2.7M CC images) with the following rule:

$$\mathbf{d}_c \leftarrow \alpha \cdot \mathbf{d}_c + \frac{(1-\alpha) \cdot \sum_{h_i=c} \mathbf{v}_i}{|\{i|h_i = c, \forall i = 1, ..., B_v\}|}, \qquad (1)$$

where h_i is the index of the found nearest neighbor in D for $\mathbf{v}_i$. $|\cdot|$ is the cardinality operator. α is the momentum coefficient set to be 0.999. The updates are run for 8 epochs.

3. Ablation Studies

Table 1 analyzes how (1) how adding attributes affect U-VB and (2) differently configured WFHs affect the downstream task performances on Flickr30K.

Adding Attributes to U-VB. One question to ask is whether solely adding attributes benefit U-VB. We tested two strategies to add the attribute tags: either by adding (`add`) the attribute tag embeddings to the object embeddings, or by appending (`append`) the attribute tag tokens along with other object tag tokens. We observe they perform on par with each other. However, to our surprise, both perform noticeably worse than U-VB, which does not consider the attribute tokens.

Our speculation is such models can be biased towards objects with specific attributes. This is because, considering the tag generator, i.e. the object detector, is trained on Visual Genome [2] where the object annotations are dominated by human-related classes, e.g. man and person, and attributes are by colors, e.g. the top five attribute classes are colors.

Modeling the limited types of object and attribute tags with WFH could have alleviated those biases given the better recall values on the retrieval tasks. While U-VB uses those tags as some fixed anchors to bridge modalities, WFH's hallucinations serve as more diverse anchors to interact with much more texts than U-VB can.

Varying Number of Layers and D's Sizes in WFH. We also test WFH with different dictionary sizes C and number of layers J. $C = 1024$ and $C = 1536$ yield comparable results while the larger $C = 3072$ shows degradation in recalls. The 2-layer WFH ($J = 2$) improves five out of six recalls (except for R@10 on IR) over the 1-layer WFH across tasks while the 3-layer option gives comparable recalls on IR, but slightly degrades on TR. These results align with our expectation – devising a larger dictionary or more layers would not necessarily lead to improvements as the visual dictionary is kept fixed over the training. It would thus be interesting to update also the visual dictionary during pre-training phase. We leave that to the future work.

4. Qualitative Studies on Cross-Modal Retrieval

We qualitatively study XMR tasks on Flickr30K in Figures 2 and 3, which include cases where either the proposed WFH model or U-VB retrieves target images or texts.

Image retrieval. As can be seen in the retrieved images in Figure 2a, the proposed model appears to be more capable of inferring occupation (e.g. army officer), details in the dress of the subject (e.g. "wide-eyed", "skull and crossbones shirt"), races (e.g. "Asian women"), compared to U-VB.

While in some cases U-VB retrieves the target images within the top-five images, our model retrieves reasonable images as well. For instance, for the query "Several peo-

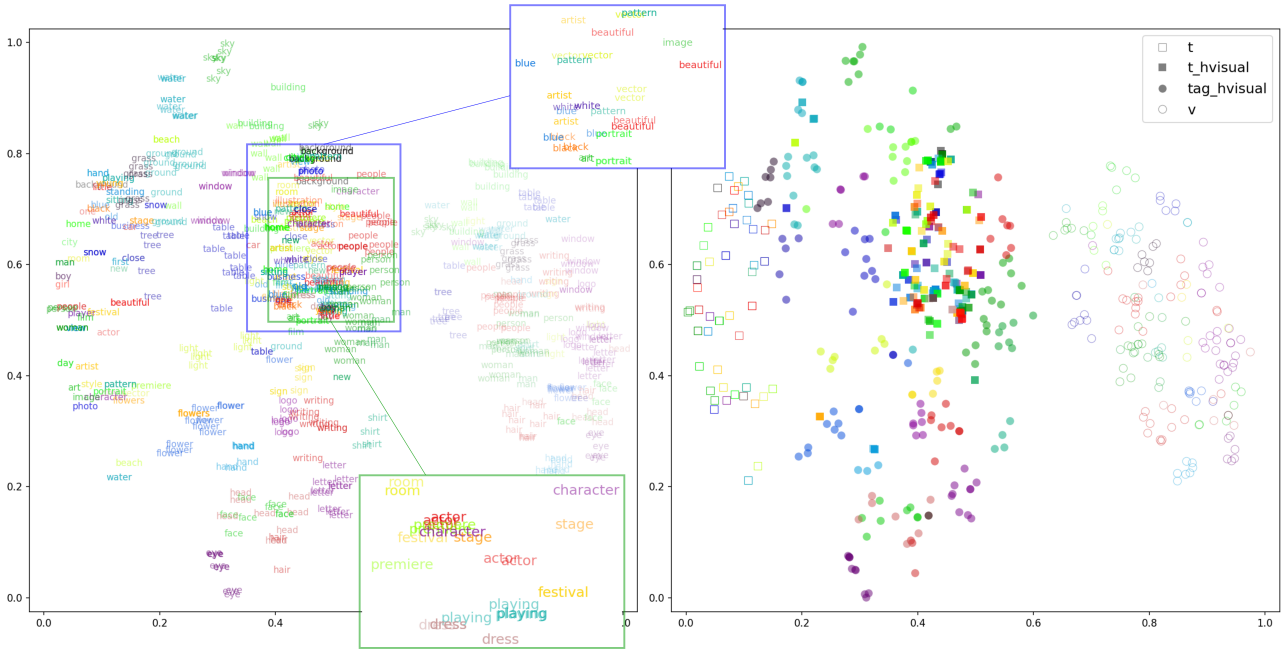


Figure 1: t-SNE plots [3] on WFH-generated representations from text tokens and object tags. Legend: "t" are text token representations, "v" are regional visual representations, "t_visual" and "tag_hvisual" are WFH-generated representations from text and object tag tokens, respectively. The zoom-in windows in blue and green display the selected `t_hvisual` samples within the respective regions. Best viewed in color.

Table 1: Comparing models on varying variables. The first two rows, `add` and `append`, are the results from the models without WFH. The rest of them are with WFH of varying configurations.

Studied variables	Text-Image Retrieval			Image-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Ways to utilize attribute tokens						
`add`	52.5	**81.3**	**88.3**	65.5	**89.7**	94.8
`append`	**52.9**	80.6	88.1	**67.0**	89.5	**95.1**
Visual dictionary size C (number of WFH layers = 1)						
$C = 1024$	55.0	**82.7**	89.8	**71.7**	**91.4**	94.8
$C = 1536$	**55.5**	82.3	**89.9**	71.4	90.9	**95.6**
$C = 3072$	54.2	81.9	89.1	69.8	90.3	94.1
Number of WFH layers J (visual dictionary size $C = 1024$)						
$J = 1$	55.0	82.7	**89.8**	71.7	91.4	94.8
$J = 2$	**56.7**	82.8	89.5	**72.5**	**91.5**	**95.6**
$J = 3$	56.4	**83.0**	89.7	70.4	90.7	95.4

ple are eating lots of food" in Figure 2b, the second image retrieved by the WFH model well capture the activities indicated in the given text.

Text retrieval. From the top retrieved sentences in Figure 3a, the proposed WFH model better recognizes, e.g. the ethnic group in the second query image, while precisely retrieving the specific object, e.g. the "ticket machine", "megaphone" and "harp" in the first, second and third images, respectively.

We next examine the cases where the WFH model does not hit the target texts in the presented cases (where U-VB does) in Figure 3b. One could still see that, for instance, for the query image listed on the top, most retrieved sentences (the first, second, third, and fifth) well describe the main subject (the man) and the activity (e.g. going into the water).

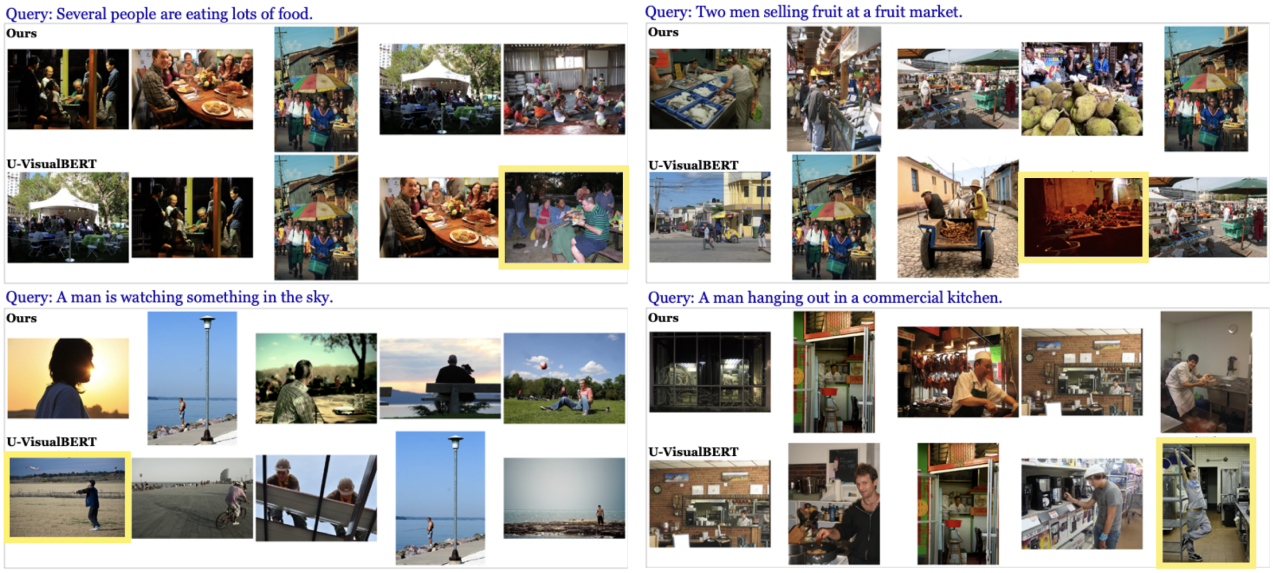
References

[1] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021.

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

(a) Cases where our proposed WFH model retrieves the target images, highlighted with the yellow frame.



(b) Cases where U-VisualBERT retrieves the target images, highlighted with the yellow frame

Figure 2: Text-to-image results where our proposed model retrieves better images. Each frame contains two rows of five images, where the first and second rows present top five images retrieved from the proposed model and U-VisualBERT, respectively.

U-VisualBERT
A man touches his back pocket as a blond woman behind the counter of a convenience store looks at him.
This man and women lost something and is looking for it.
A man is leaning over and pulling something out of a bag.
A man in glasses and an Asian woman are seated across from each other in a subway train.
A woman in her early fifties walks by herself in the subway, another woman is not far behind.

Ours
A man working on a ticket machine as two women stand near.
A woman wearing rolled up jeans and a black shirt is walking past a bus carrying a shopping bag.
A woman looks on as a man with folded arms is talking.
A man in glasses and an Asian woman are seated across from each other in a subway train.
A guy and a girl having a conversation.

U-VisualBERT
A crowd is assembled in a street.
A child stands outside with a crowd of adults in white.
Three young adults talk in a crowd of people, the woman looks upset.
A group of young people stand around waiting for something.
These people are walking in a crowd of people.

Ours
A woman in a white shirt and hat speaks to a large crowd of men and women using a megaphone.
An African group of men, women, and babies pose in a field with large hills in the background.
Africans at an organized event. (target)
A teenage boy with a white headband looking right, a crowd in the background.
A crowd is assembled in a street.

U-VisualBERT
Man plays a strange looking string instrument.
A person is playing an unique instrument.
A man wearing leather pants, a leather vest and a cowboy hat is standing here.
A man is leaning over and pulling something out of a bag.
A man holds out a wire basket as he prepares to fill it.

Ours
A woman playing a harp.
A person is playing an unique instrument.
A female harp player peers through the middle of her instrument while performing. (target)
Man plays a strange looking string instrument.
Where is the rest of his racket?

U-VisualBERT
While working a woman makes a peace sign and smiles so big it makes her eyes close.
An asian girl is flashing the peace-sign.
A very young boy stares ahead as he is biting onto a small object.
One little boy working hard in his martial arts class.
An Asian woman in a white shirt working with machinery giving the peace sign.

Ours
An woman wearing a light-blue shirt adjusts her hair while sitting at a table.
A young girl is sitting backwards on a chair, looking over the back of it.
An asian woman demonstrating cooking with margarine in a public place
A young girl sitting on a wooden chair.
A girl sitting on a chair backwards.

U-VisualBERT
Three people sitting in a large bike basket.
A man demonstrates how to ride a bicycle with only the back wheel on the ground.
A man assembles a bicycle for a little girl.
Man laying on the ground looking at a boy.
A guy fixing a bike for a young girl.

Ours
A man working in the ground in the dirt looking away.
Two boys try to attach fruits to a bicycle on a dirt road.
Two young men getting ready to move fruit on their bike.
A man assembles a bicycle for a little girl.
A man demonstrates how to ride a bicycle with only the back wheel on the ground.

(a) Cases where the proposed WFH model retrieves the target sentences, which are highlighted in blue.

U-VisualBERT
A man hangs upside in the air over the water.
A person upside down on a water board.
A person is swinging by a cord over the ocean.
A man wearing bathing trunks is parasailing in the water.
The feet are shown of a person who just fell off of their surfboard.

Ours
A man in a harness lands in the water.
A man on a tag line going into the water.
A man on a zip line being propelled through the water.
A man wearing bathing trunks is parasailing in the water.
A person is swinging by a cord over the ocean.

U-VisualBERT
A woman in a sweater is laughing.
A person in red with their hand raised and fingers stressed.
A woman in a red shirt is raising her arm to the crowd below.
A young female student performing a downward kick to break a board held by her Karate instructor.
Possibly a preschool teacher trying to explain something to her class, the little mouse could care less.

Ours
A person in red with their hand raised and fingers stressed.
A woman in a pink sweatshirts holds a bouquet of balloons while sitting on a folding chair.
A person in a red jacket with black pants holding rainbow ribbons.
A young girl has lifted her friend and is carrying her in her arms.
A girl kicking a stick that a man is holding in tae kwon do class.

U-VisualBERT
A man wearing all white (including a bandanna) cooking something and making a huge flame.
A man watches a pan catch fire while cooking in the kitchen.
The man in a Japanese cooking suit is preparing a meal for two people.
A man in a chef's uniform holding a large skillet over a stove, with fire coming out of the skillet.
An Asian man in white working in a kitchen.

Ours
A man wearing all white (including a bandanna) cooking something and making a huge flame.
A man watches a pan catch fire while cooking in the kitchen.
A man in a chef's uniform holding a large skillet over a stove, with fire coming out of the skillet.
A man with an apron and hat cooking.
A person cooking on a fire while a small child watches.

U-VisualBERT
A chinese man sitting down waiting for customers.
People sit in an Asian restaurant.
A man in a checked shirt is sitting at a table looking back at a group of people behind him.
A man sits in an outdoor cafe finishing a meal.
People are gathered around the table filled with food.

Ours
Several people are eating lots of food.
Several groups of seniors eat food at different tables.
The picture shows several people eating at outdoor tables.
People sit in an Asian restaurant.
A man eating at an outdoor restaurant

U-VisualBERT
A little girl finds joy in playing in a big, sloppy area of mud!
A naked little girl plays in a mud puddle.
A little girl running at on the shore of a beach.
Little girl standing in sand at the beach.
Little girl in arm floaties exploring the coast line.

Ours
Little girl standing in sand at the beach.
A little girl running at on the shore of a beach.
A little girl finds joy in playing in a big, sloppy area of mud!
A naked little girl plays in a mud puddle.
A young child is running along a beach.

(b) Cases where U-VisualBERT retrieves the target sentences, which are highlighted in blue.

Figure 3: Image-to-text results. Each frame shows top five captions retrieved by models. The first (upper) and the second (lower) frames display results from U-VisualBERT and the proposed WFH model, respectively.