Reviewer 1

> *"The results (Table 2) show similar accuracies for all attention models."*

Note that the VQA results in Table 2 with continuous attention use fewer basis functions than discrete regions. Although the accuracies are similar, the unimodal attention suggests better interpretability (as noted by R2 and R3).

> *"It would help to have a short algorithm describing how to implement the forward and backward passes efficiently."*

Good idea, we will add this to the camera-ready version.

> *"Line 134: "a condition (...) is g being strongly convex". Is this a necessary or a sufficient condition?"*

Sufficient; we will clarify and follow the suggestions (move the beta-escort definition to the main text and fix typos).

Reviewer 2

Thanks for the positive comments and for pointing out the work of Cordonnier et al. (2020). We will add a citation.

> *"Have you experimented with other ways such as linear interpolation (1D) or bilinear (2D)?"*

We chose ridge regression as it enables a closed-form solution expressed linearly in terms of the basis functions (Eq. 15) and matrix G can be precomputed, leading to a fast implementation. Note that the proximity of tokens/pixels is taken into account (the basis vectors $\psi(t_\ell)$ forming F are located at each token/pixel). We haven't tried linear interpolation, but this is an interesting suggestion (although it might make attention computation more challenging).

> *"Does it remove the need for additive positional encoding?"*

Very good point; this is indeed one advantage of our approach – by converting the input to a function on a predefined continuous space, it encodes "positions" implicitly in a natural way, not requiring explicit positional encoding.

Reviewer 3

> *"The proposed method's application on VQA is limited to grid feature."*

Actually, our method can handle BUTD features too: it suffices to let the t_ℓ coordinates in the multivariate regression (Eq. 15) be placed on these regions instead of on a grid. However, we opted not to rely on an external object detector, in order to check if continuous attention has the ability to detect relevant objects on its own (see ellipses in Fig. 3). However, for a high-level vision system, combining our method with BUTD is an interesting idea.

> *"Why model text as continuous inputs? Text are naturally discrete tokens."*

We agree text is fundamentally a discrete sequence of symbols. However, when processing long documents or attending to snippets, modelling it as a continuous signal may be advantageous, due to smoothness and independence of length.

> *"The proposed method assumes attention probability is single mode. Is this a reasonable assumption?"*

Good point. Unimodal attention is useful to focus on a single object or text segment of varying size, avoiding "fragmenting" attention probability; however, in some applications, multimodal attention may be preferable. Our method can be extended to multiple modes via a suitable choice of $\phi(t)$ (e.g., a mixture of Gaussians), but this will require numeric integration for attention computation. A simpler strategy (see lines 258-260) is to use multi-head or sequential attention.

> *"Can it be applied to deep transformer models with multi-head attention?"*

Great question. We have ongoing work applying this to transformer models (but out of scope for this paper). Briefly, the computation cost is $O(N)$ for each attention head (against $O(L)$ in the discrete case) where $N \ll L$ is the number of basis functions, plus an extra $O(NL)$ cost in the first layer to perform the multivariate regression on L tokens.

> *"Is there any intuition (...) how the continuous attention can improve the accuracy of downstream tasks?"*

In general, continuous attention can make it easier to attend to large spaces with different resolution levels, with a fixed number of Gaussian RBFs with several variances. It can also lead to more focused attention (the VQA experiments suggest this) and better control of time steps with continuous data streams (*e.g.*, irregularly sampled time series). We haven't explored all these directions, but we believe these are promising areas of future research.

Reviewer 4

Please check our answers to R1 and R3 above (the answer is yes to positional encodings).